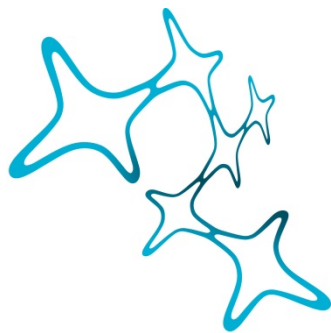


---

# Metacognition of value-based decisions

Oriane Armand

---



Graduate School of  
Systemic Neurosciences

LMU Munich



Dissertation at the  
Graduate School of Systemic Neurosciences  
Ludwig-Maximilians-Universität München

Munich, 24<sup>th</sup> of July 2023



# *Reviewers*

## *First Reviewer:*

Prof. Dr. Ophelia Deroy  
Chair of Philosophy of Mind  
Ludwig–Maximilians–Universität München  
Faculty of Philosophy, Philosophy of Science and the Study of Religion

## *Second Reviewer:*

Prof. Dr. Stephan Sellmaier  
Head of Research Center for Neurophilosophy and Ethics of Neuroscience  
Ludwig–Maximilians–Universität München  
Faculty of Philosophy, Philosophy of Science and the Study of Religion

## *External Reviewer:*

Prof. Dr. Simone Schütz-Bosbach  
Professor of Experimental Neuro-Cognitive Psychology  
Ludwig–Maximilians–Universität München  
Department of Psychology

Date of Submission:

24<sup>th</sup> July 2023

Date of Defense:

8<sup>th</sup> February 2024

# *Acknowledgment*

*“If I have seen further it is by standing on the shoulders of Giants.”  
Isaac Newton, 1675*

I remember my first interview with Prof. Dr. Ophelia Deroy as if it was yesterday. I recall the surprise when before discussing my scientific skills and the project at hand she preferred to openly ask about my academic interests and aspirations. If great leaders do not impose but provide support and guidance then I shall be eternally grateful to have had one as a supervisor for my PhD. Besides providing me with the unique possibility as a scientist to design a PhD project of my own choosing, Prof. Deroy provided me with the intellectual freedom and rigour to turn this ambition into a reality. Last but not least, I am grateful to her to have provided me with, further than a remarkable academic environment, a home away from home, the CVBe where we all feel like we belong together.

I wish to sincerely thank Prof. Dr. Stephan Sellmaier for his continued cheerful support and optimism together with a very helpful supervision in finishing my thesis. Dr Sellmaier's wisdom and expertise when it comes to academia were essential keys to bring the present thesis together not merely as a list of scientific and academic chapters but rather as a structured and meaningful whole which I aspired to complete when joining this interdisciplinary program. I am extremely grateful to Dr. Benedetto de Martino for having provided me with guidance during my PhD. His scientific expertise and philosophical mind were endless sources of inspiration to

my work and his sense of humour and openness very precious uplifts on the road. Lastly, I wish to thank Prof. Dr. Thomas Schenk for his guidance and his views in both the field of cognitive neuroscience and in the journey that is a PhD.

I would like to thank the members of our interdisciplinary group for their friendship and academic support: Sofia Bonicalzi for including me in her EEG projects and introducing me to the field of freewill; Merle Fairhurst for her ever kind and optimistic guidance; Nora Heinzelman for her views on action theory, Justin Sulik for this irreplaceable help with my data analyses and Jurgis Karpus for his knowledge in economics. I also feel very lucky to have shared the journey of being an interdisciplinary PhD student together with Anita Keshmirian, Sofiiia Rappe, Mark Wulff Cartensen, Louis Longin, Harry Waterstone and mainly my desk neighbour Lucas Battich who's rare patience helped to appease my seemingly endless curiosity for philosophy.

I am also grateful to have been included in the online meetings of Benedetto de Martino where I received feedback on my work and was inspired by the projects of Aurelio Cortese, Mariana Zurita and Pradyumna Sepulveda. I also wish to thank for their particularly helpful feedback on my work Bahador Bahrami, Joaquín Navajas, Mehdi Keramati, Marion Rouault and Nicholas Shea. I also wish to thank other PhD students of the program of Neurophilosophy for their philosophical feedback on my work. Also I am sincerely thankful to the Graduate School for Neurosciences for opening me the door of their program until then reserved to philosophers and thereby providing me with the unique opportunity to design an empirical PhD on metacognition as shaped through the eyes of philosophy.

Last but not least I would like to sincerely thank my parents and family for their continued support along the journey and so despite the long silences. I am also very grateful to my baby niece Cléo who's joy and interest in horses are the most indescribable blessings.

*To Hayao Miyazaki.*

*Who's masterpieces have inspired an entire generation of little girls to thrive to become their own moral heroes.*



# *Summary*

*“If knowledge is power, then self-knowledge is empowerment.”*

*Pedro Gaspar Fernandes*

Heroes inspire us with their ability to remain true to their own values, even in complex tasks involving risk, instinct, or social norms. Resembling a super power, this autonomy in their behaviour can be pinned down to a consistent coherence between their decisions and their intentions. We refer to such life-like decisions that rely essentially on subjective preference as value-based choices. Metacognition is thought as the ability to monitor and control one’s own decisions, and thereby is commonly agreed to provide humans with a unique capacity to ensure this coherence even in complex tasks. However, while metacognition has been demonstrated to monitor simple hedonic decisions with confidence reports, it remains unclear whether metacognition can track the coherence of other types of value-based decisions, such as moral ones. Additionally, it is unclear whether confidence reports can inform behaviour to ensure this coherence and ultimately contribute to making a hero. To better understand whether the core function of metacognition truly is to track and ensure coherence in such life-like value-based decisions, this thesis asks:

**What does accounting for the subjective value of decisions contribute to our understanding of the function and computation of metacognition?**

In other words, how does metacognition use subjective value to monitor and optimise behavioural coherence, ultimately creating driven agents such as virtuous heroes or even super villains. This thesis presents novel empirical and conceptual work that helps define the different facets of this ubiquitous metacognitive monitoring in these life-like value-based decisions.



Part 1 of the thesis establishes a conceptual framework for metacognition of value-based decisions. Chapter 1 provides a general introduction to how value-based metacognition monitors the coherence of decisions by accounting for subjective preferences. In chapter 2, we present how both philosophy and cognitive neuroscience define a landscape of procedural metacognition. Merging this functional approach with bounded rationality, we approximate metacognition to a reliability thermostat where a desired degree of accuracy is maintained by adjusting the level of effort applied based on various monitoring signals. In chapter 3, we introduce a novel architectural model of metacognition where the value-based side of the monitoring process appears as an inherent facet of its computation, informing the agent about contextual reliabilities and supporting learning.

Arguing that value-based monitoring is an ubiquitous part of the metacognitive process, Part 2 of the thesis tests this hypothesis. In chapter 4, we present preliminary data on decision making with limited knowledge, and suggest that reflecting upon one's own decision (metacognition) does not rely on cues for accuracy but on heuristics seemingly reflecting other's decisions (theory of mind). In chapter 5, we demonstrate that participants can track the coherence between their decisions and their preference in moral (as well as hedonic) tasks and that these confidence judgments predict choice repetition. In chapter 6, we go beyond the norm of coherence and suggest with preliminary data that participants can have subjective metacognitive profiles that are consistent across value domains by either tracking choice optimality or satisfiability.

Lastly, chapter 7 concludes the thesis by discussing its contribution to ongoing research: it positions value-based metacognition as a pioneer area, opening pathways for real-life applications in autonomy (as a social agent) and in metacognitive enhancements and therapies to boost or restore cognitive functions. By framing metacognition in terms of subjective value and coherence, this thesis provides a foundation for understanding how metacognition can enhance human autonomy and responsibility.

# Contents

Acknowledgments	iv
Summary	vi

## **Part I: A theoretical framework.**

<b>1. General Introduction</b>	<b>1</b>
1.1 A norm of behavioural coherence	3
1.2 Operationalizing insight	9
1.3 Thesis Synopsys	12
Bridge: From introduction to landscape	17
<b>2. The Metacognitive Landscape, a multi-disciplinary challenge.</b>	<b>19</b>
2.1. Introduction	21
2.2 Philosophy of metacognition	23
2.2.1 A meta level landscape	23
2.2.2 Social Origins	35
2.2.3 Responsibility	38
2.2.4. Non-human intelligence	42
2.2.5. Intermediary conclusion	43
2.3. Cognitive Neuroscience of metacognition	45
2.3.1 Computation of confidence	46
2.3.2. Functions of confidence	59
2.3.3. Subjective factors of metacognitive ability	65
2.4. Conclusion	68
Bridge: From landscape to learning	79
<b>3. Inferential Metacognition of Perceptual and Value-based decisions</b>	<b>81</b>
3.1 Introduction	84
3.2 Inferential metacognition	86
3.2.1 Multi-dimensional monitoring	86
3.2.2. The structure of cognitive inference	87
3.2.3. Monitoring the reliability of different sources	89
3.3. Local confidence and learning	90

3.3.1 Complementary monitoring of PC and VC	90
3.3.2. PC and VC in learning	98
3.3.3. Local control	100
3.4. Global confidence	101
3.4.1 Monitoring Global Confidence	101
3.4.2 Global confidence	106
3.5. Conclusion	108
Bridge: From inference to confidence	117

## **Part II: Empirical studies.**

<b>4. Confidence in art: the consensual illusion of accuracy.</b>	119
4.1 Introduction	121
4.2. Methods	124
4.3. Results	127
4.4. Discussion	133
4.5. Supplementary material	141
Bridge: From heuristics to moral value	149
<b>5. Confidence monitors and predicts moral decisions.</b>	151
5.1 Introduction	152
5.2. Methods	156
5.3. Results	158
5.4. Discussion	163
5.5. Conclusion	167
5.6. Supplementary Figures	176
5.7. Supplementary Material	182
Bridge: From moral to general	185
<b>6. Beyond the rational monitoring of value-based decisions.</b>	187
6.1 Introduction	188
6.2. Methods	189
6.3. Results	191
6.4. Discussion	198
5.5. Conclusion	200

5.6. Supplementary Figures	204
<b>7. General Discussion</b>	211
7.1 Our results support a value-based model of metacognition	213
7.2 Remaining questions for the model	217
7.3 Open discussion and implications of the model	220
7.4 General conclusion	230
Annexe: Interdisciplinary Glossary	233
Curriculum Vitae	239
List of Publications	242





# Chapter 1

## General Introduction

- “On the charge of the theft of the Goya portrait of the Duke of Wellington, do you find the defendant guilty or not guilty?”
- Not guilty.”

Our society assumes that civilians are coherent agents whose acts and intentions go hand in hand. Therefore, if an agent acts wrong, his intentions presumably are wrong and are to be legally reprimanded for the good of the social group. Equally, if an act is heroic then the agent's intentions presumably are also those of a hero who should be praised. If the good working of a society relies on this parallel between the agent's intentions and actions, then one can ask what is the human ability that ensures this coherence? In the recent movie *The Duke* (2020), this foundational assumption of our legal system is highlighted when the “modern Robin Hood” Kepton Burton is cleared of his charges for “stealing” from the National Gallery. The essence of the case relies on the consistent demonstration of the man's good intentions despite his seemingly criminal act. There, this ability to remain true to his good intention despite these actions seeming wrong defines the making of a hero. The movie narrates this situation with several letters sent by the protagonist to the authorities where he argues for the legitimacy of his act in the light of his intentions:

“The act is an attempt to pick the pockets of those who love art more than charity. My noble intention is merely to raise £140 000 to fund a charitable cause of my own choosing. If the fund is raised, the picture will be handed back.”

These explicit testimonials of Mr Burton's assessment of his action as legitimate in the light of his good intentions had two main implications. First, communicating his optimism in the outcome of his act comforted Mr Burton himself that his action was worth sticking to, presumably consolidating his heroic resilience to act in accordance to his intentions. Second and most importantly here, this explicit testimonial provides a concrete proof of his belief in the goodness of his act, which enabled his lawyer to remind the jury about a pillar of the British law: the assumption that agents act in coherence with their intentions and that, therefore, well-intended civilians should generally not be punished for the rare occasions when their acts result in unintended harm. The lawyer reminds this fundament of the legal system as follows:

“Nothing is a crime in this country unless it is expressly forbidden by law. If your neighbour borrows your lawn mower and does not return it for months: it is frustrating, it is annoying, but it is not theft because he had no intentions of permanently depriving you of it. Kempton Burton is your neighbour, he is not a thief. He borrowed your Goya to try and do a bit of good in this world.”

Indeed, by distinguishing actions (*actus reus*) from intentions (*mens rea*), the British law therefore stipulates that for most crimes, an agent can only be charged as guilty if the harm caused was intended. In other words, our penal system is ready to overlook some unfortunate outcomes as long as they were performed with good intentions because it assumes that citizens are autonomous and responsible and thereby, by acting coherently with their good intentions, would not require punishment. But in such circumstances where what is good and bad is hard to dissociate or drown in risk and uncertainty, what is the underlying skill an agent must possess to remain autonomous and act in coherence with her intention? Would such an ability that provides autonomy and coherence be the making of a moral hero, or even a super villain? Is it to some degree ubiquitous to all humans and to all circumstances? Could such a gift also be a curse?



In this thesis, we focus on this ability of Mr Burton that enables him to carry out this heroic action and also to be recognised as such by society instead of being incriminated: the ability to explicitly monitor whether his own actions are coherent with his personal values. More specifically, in this thesis we take the perspective that the ability to monitor one's actions intrinsically relies on subjective values, and therefore, looking through this lens, ask how this perspective can contribute to better understanding this explicit monitoring (i.e. metacognition). In other words, we ask:

**What does accounting for the value of decisions contribute to our understanding of the function and computation of metacognition?**

In this short introduction, we start by defining coherent behaviour both for its philosophical assumptions and its operationalisation in economics. Then, we present the essential concepts and operationalisation of the study of decision monitoring (i.e. metacognition) and open on its subsequent implications on coherent behaviour. Lastly we present the synopsis of the thesis chapters with their respective contribution to this multi-facet question.

## **1. A norm of behavioural coherence**

### **1.1. Philosophy of autonomous coherence**

What are the common denominators of a coherent action? Action theory is concerned with defining the driving forces of an agent's actions. In its simplest form, Davidson and Kant propose such conditions for actions. Davidson (Davidson, 1982) defines that an agent's beliefs (*e.g.* the box contains a marshmallow) and desires (*e.g.* hunger) jointly form the reason for her action (*e.g.* takes the box). These "reasons" therefore form the causal precursor that results in the action. Kant contrasts two types of reasoning in regards to rationality: theoretical reasoning forms a set of coherent beliefs that aims at truth whereas pragmatic reasoning constitutes a coherent set of strategies to fulfil one's ends. Both philosophers therefore agree on the view that having a set of coherent motives result in a coherent behaviour. Adding a level of complexity to these models, Frankfurt (Frankfurt, 1971) proposes that a hierarchical structure of such desires (or motives) provides a reliable structure for coherent behaviour. He suggests that despite local desires (*e.g.* wanting to eat the

snack), higher order desires guide these latter (e.g. wanting not to desire unhealthy food). He defines that such incoherent desires can end up by making us “choose what we do not want and want what we do not choose”, and that incoherent behaviour can therefore arise from conflict between these different levels of desires. In other words, Frankfurt’s hierarchy of desires can be seen as the modulation of lower desires by other desires such as driven by long-term goals. This distinction between first and higher order desires can be illustrated in our previous example. For the case of Kepton Burton, the direct consequences of taking the painting from the national Gallery is to deprive the nation from its value and should therefore be acknowledged by the agent and the jury as against (i.e. incoherent with) what is expected from a good citizen. However, because this modern Robin Hood had from the start this higher desire to shed light on economic injustice and was confident in the goodness of this intention, from the start he therefore communicated coherently to the authority that this apparent act of theft was actually nothing but a momentary borrow for the greater good. In the eyes of the law, this coherent and consistent communication his higher order desired reframed the incriminating act as not guilty. Similarly, action theory suggests that an action is coherent if it follows a (hierarchically) organised set of desires which the agent is consciously aware of. The given example therefore both illustrates the concept of coherent behaviour and highlights its implication in the legal system on which relies our society. In the following part we discuss how this philosophical concept of decision coherence as aligning with subjective intention can be modelled and operationalised to be studied empirically.

## **1.2. Economics of rational coherence**

### **1.2.1. Normative models**

Classical economics founded normative decision models that define the rational or best decision an agent should make. The concept of utility provides a common ground to model the agents’ reasons for action as a common currency of values such as pleasure, happiness and welfare (Mill, 1863). In this utilitarian framework, agents are expected to choose therefore in a fair manner considering the value for the greater good and are also assumed to have complete knowledge with no limitation

in their ability to compute the decision's outcomes. Building on this framework, Smith proposed to restrain the concept of utility to a self-centred accumulation of wealth which provided the foundation for the development of the economic models of a rational *homo economicus* (A. Smith, 1986). For instance between A- obtaining \$10 tomorrow and B- obtaining \$20 next week, *homo economicus* would rationally choose the latter alternative to maximise his gain. At the heart of modern economics, the utility function portrays the relative preferences of the agent as follows:  $u(A) < u(B) = \$10_{\text{tomorrow}} < \$20_{\text{next week}}$ . This function therefore defines (with the above assumptions) the normative behaviour or best decision an agent should make.

Most popular in decision theory, Expected Utility introduces the notion of uncertainty in the picture. In this framework, rational behaviour is defined by maximising (non-exclusively egoistic) utility defined by the linear combination of utilities and their probability of occurrence. Often relying on lotteries, an illustration of this normative model predicts that an agent would prefer a lottery A with a 80% chance of obtaining \$10 over a lottery B with a 20% chance of obtaining \$20. The utility function predicting the agent's rational behaviour would describe the preference  $p(A)*u(A) > p(B)*u(B)$ . These normative models therefore provide a norm of correctness to study when an agent would deviate from optimal behaviour. In the following part we define however more descriptive models that account for human limitations to these models.

### 1.2.2. Descriptive models

In contrast to normative models which describe what rational agents ought to do to maximise outcome, descriptive models thrive at predicting what real agents are actually most likely to choose. Detaching from these classic normative models, Prospect Theory founded neoclassical economics by using cognitive psychology to define common sub-optimal biases. With his central loss-aversion principle or non-linear probability weighting, Kahneman highlights the consistent contribution of sub-optimal heuristics in human decision making. With its non-normal curved sensitivity, the concept of loss aversion predicts (depending subjective aversion) an agent could prefer a lottery A- with 100% chance of gaining \$9 over a lottery B- with

a 95% chance of gaining \$10. Accounting for risk aversion, prospect theory therefore provides an explanation for the observed preference  $1 * \$9 > 0.95 * \$10$ .

To further account for real life decisions, bounded rationality takes into account the agents limitations. According to this framework, agents do not aim at maximising outcome but at making efficient decisions by maximising a cost-benefit ratio: if obtaining an additional \$10 requires waiting for a week, then the cost of waiting might be greater than the actual gain which results in the preference  $\$10_{\text{tomorrow}} > \$20_{\text{next week}}$ . Together with Prospect Theory, Bounded Rationality presents some violation of rational axioms by presenting consistent decisions biases and fallacies. Unlike the normative models that describe how an ideal observer would rely on a “normal” value space to make decision, descriptive models instead aim at capturing how cognition distorts and shapes a subjective value space and in turn guides decision.

To simplify these complex relations and expectations, here we reduce the norm of behaviour to the concept of coherence which predicts that that preferences which are explicitly expressed by participants would predict their decisions. In other words, we expect that the conscious preferences and their ordering take into account the participant’s many cognitive limitations and distortions of value and are therefore able to reliably predict decisions.

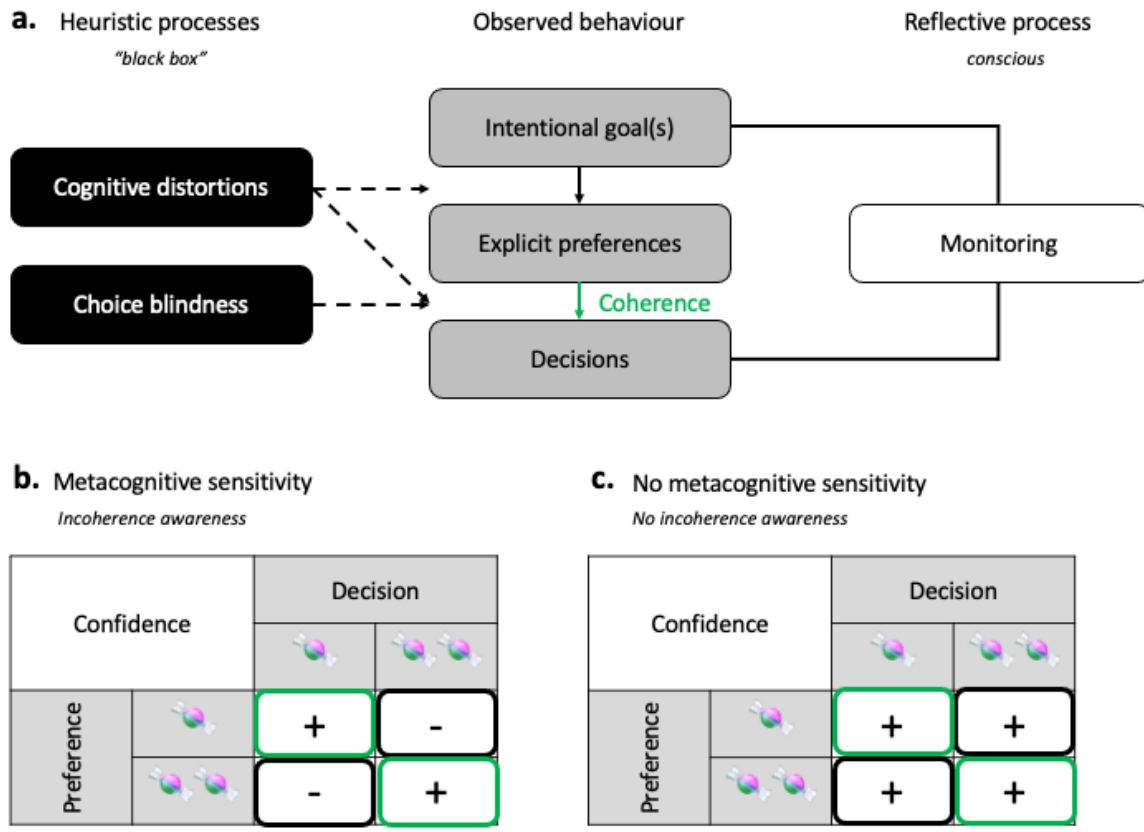
### 1.2.3. Eliciting preferences

In behavioural economics, preferences are defined as the ordered ranking of the subjective utilities for a set of items. To obtain this subjective ranking and subsequently study preferential decision making, the gold standard method is to ask the participant to report her “willingness to pay” ( $WTP(A) \approx u(A)$ ) for each item (Chapman et al., 2017). Often done in lotteries to study subjective risk aversion, the method consists in asking each participant the finite amount of money she would invest to play each lottery in a set (e.g.  $WTP(\text{lottery B: } 0.95 * \$10) = \$0.9$ ). This method has also been extended to other economical decisions with real life items such as snacks to represent utility of hedonic value (e.g.  $u(\text{chocolate ice cream}) > u(\text{vanilla})$

ice cream)). Within this framework, axioms of rationality define normative behaviour within a finite set of item by predicting:

- transitivity: if  $u(\text{chocolate}) > u(\text{strawberry})$  and  $u(\text{strawberry}) > u(\text{vanilla})$ , then  $u(\text{chocolate}) > u(\text{vanilla})$
- completeness: agent has either strict preference:  $u(\text{chocolate}) > u(\text{vanilla})$ ; or indifference:  $u(\text{chocolate}) = u(\text{vanilla})$ ; or weak preference:  $u(\text{chocolate}) \geq u(\text{vanilla})$
- non-satiation:  $u(\text{lots of good}) > u(\text{little good})$  and  $u(\text{little bad}) > u(\text{lots of bad})$

This set of axioms can be criticised on two points. First it describes normative behaviour of optimality which, as previously discussed, might be over ruled by descriptive model that take into account the value distortions of human cognition. Secondly, this ordinal system considers preferences based on their ranking uniquely and can be refined in the form of a cardinal system that weights preferences on a continuous measure of utility. Aligning with the cognitive considerations of descriptive models, accounting for this continuous measure of utility enable to account for decision difficulty, an important aspect to account for when studying human behaviour. Now that we defined the normal and descriptive models of decisions and their operationalisations, the following part defines how, from the subjective point of view, an agent evaluate her own decisions as successful or erroneous.



**Figure 1: Coherence of the decision-making process: between heuristics and monitoring.** **a.** Schematic representation of some of the main cognitive processes influencing the coherence of decision-making (as inspired by the dichotomic view of Kahneman’s Dual Process Theory). As in behavioural economics that study cognition through observable behaviours (grey boxes); here we assume that: 1- an intentional goal has to be maximised without conflicting with other goals, 2- that this leads to the explicit elicitation of an utility function for a given set of items and that 3- decisions are made in coherence with these preferences. On the left side we present some heuristic processes (black boxes) that may affect the decision-making process with (mainly unconscious) shortcuts and lead to incoherent decisions. The dotted arrows represent a likelihood that these heuristics might affect the decision process and therefore affect or not the coherence between explicit preferences and decisions. On the right side, the reflective process (white box) monitors the fit between the goal maximisation and the decisions made. Reflective processes relying on higher-order cognition such as metacognition are here assimilated to the the computation of confidence levels in value-based decisions. Altogether this diagram highlights the possible gap between the processes informing the decision and the

process monitoring it, leading to eventual failure of monitoring. **b-c.** Illustration of the monitoring sensitivity or failure. The scenario presents the agent with a two-by-two scenario where an agent can either choose one or two sweets and prefers one or the other depending on the goal to respectively minimise or maximise sugar intake. In each table, the observable behaviour (grey boxes) presents for each preference (row) the agent's decision (column) as either coherent (green frame) or incoherent (black frame). The tables differ in the monitoring system's sensitivity (left table) or failure (right table). Here we equate the monitoring process (white boxes) to explicit metacognition with reports of confidence as either high (+) or low (-). A metacognitively sensitive agent discriminates coherent (green frame) from incoherent (black frame) decisions with respectively high and low confidence. Metacognitive sensitivity therefore provides the agent with the awareness and ability to communicate his goal independently of his choice and its coherence. Together these diagrams highlights the possible distance or connection between the failures of the cognitive and metacognitive processes, respectively (black box heuristic) incoherence and (white reflective) lack of sensitivity.

## **2. Operationalizing reflection**

### **2.1. History of reflection**

The study of behaviour from the first person perspective can be traced back to Freud who, through psychoanalysis, aimed at identifying the driving forces behind his patients' inadequate behaviours. Nonetheless, based on subjective storytelling, his method was not fit to develop an objective model for reflective thinking. Later, by focussing on the study of memory, the research on cognitive development and learning provided the first building blocks for a systematic measure of self-monitoring. In 1967, Hart published the first correlational method that captured the fit between subjective evaluation of decisions and their objective accuracy. In the decade that followed, Flavell described the term of metamemory (1977) as "monitoring and knowledge of storage and retrieval operations" and soon after the term metacognition (1979, and later Nelson & Narens, 1990) as relating to the "monitoring and control of cognition".

## **2.2. Defining metacognitive monitoring**

Amongst the multiple facets of metacognition, in cognitive neuroscience, the term mainly refers to its procedural and explicit form. First, as embedded in the field of epistemic metamemory, two types of metacognition were distinguished: declarative metacognitive knowledge (i.e. the quality of one's knowledge about cognitive processes and strategies) and procedural metacognition (i.e. the self-evaluation of decisions for adjusting behaviour). Secondly, in contrast with its implicit forms of uncertainty such as observed in slow response time or reduced investment (e.g. waiting for reward), explicit forms of certainty can be expressed verbally such as with levels of confidence in having chosen the correct option. This ability to explicitly communicate uncertainty as confidence levels is believed to have evolved to improve collaboration in a group and to provide a better ability to adjust decisions (Heyes et al., 2020; Shea et al., 2014). In this thesis we focus on the ability to monitor one's choices with confidence levels as a procedural and explicit form of metacognition. From trial to trial, these levels of confidence can therefore be studied for both the factors that influence them and the subsequent effect they themselves have on decision adjustments.

## **2.3. Measures of accuracy**

In the laboratory, metacognitive monitoring can be studied for its ability to discriminate the objective correctness of decisions subjective reports. Independently from the normative or descriptive models of decision, confidence levels therefore provide a window into the subjective monitoring of one's own decisions quality. The reliability of a decision (or its correctness) is defined by how well the decision fits with the decision rule and therefore depends on the task. In perceptual tasks, a decision is correct if the participant accurately discriminates between two perceptual signals in respect to an explicitly instructed decision rule (e.g. the dish containing the most marshmallows is identified to be the left rather than the right one). In value-based decisions, a choice is correct if it is coherent with the participant's preferences (e.g. the preference for more rather than less marshmallows) (Hoffmann, 2016). These value-based tasks can be studied in two ways: in preferential tasks participants are presented with familiar items for which



they elicit their preferences and chose accordingly (e.g. sets of defined lotteries or of familiar snacks); in reinforcement learning, participants are presented unknown items and reveal through their decisions how their preferences evolve through feedback from the environment (e.g. two armed bandit task).

Independently from the decision rule on which participants base their decisions the accuracy of their confidence levels in this regards can take different perspectives. The concept of metacognitive sensitivity refers to the ability to discriminate correct from incorrect decisions with respectively high and low confidence levels. Independently from this latter, the concept of metacognitive bias refers to the general tendency to over or under estimate one's performance. Lastly, the concept of metacognitive efficacy related to the ratio between the participant's metacognitive sensitivity and cognitive sensitivity (e.g. performance) at a tasks. By studying how different tasks and cues or participants population affect metacognitive accuracy, research gains insight into the building blocks of confidence levels (Fleming & Lau, 2014).

#### **2.4. Models of confidence**

While descriptive models of confidence define how confidence relate to objective correctness of decisions, the research aims at identifying the cues from the decision making process on which this inference relies. The two main predictors of confidence as demonstrated by perceptual decision making are signal strength and noise. Respectively these two parameter define the difficulty of the choice such as by representing the contrast between two items (1 vs. 10 marshmallows is a greater signal strength than 2 vs. 3) and the access to the evidence (e.g. if the marshmallows are right in front of us the signal is clearer than if they are at 20 meters from us). Choice difficulty predicts both correctness and confidence in perceptual and value-based decision making (Sepulveda et al., 2020). Beside the signal relating to the decision rule, other predictors were demonstrated to define confidence levels such as the time taken to make the response (response time) or the confidence in the previous trial (Shekhar & Rahnev, 2020). Understanding the evidence on which relies confidence enables researchers to identify the tasks in which agents might have a limited ability to monitor the correctness of their decisions. In this thesis we

investigate the working of metacognition in value-based decisions and discuss a framework in which this monitoring appears ubiquitous to real life decisions. More specifically, in chapters 4, 5 and 6 we present new empirical data that defines how confidence levels are shaped both in decisions with limited knowledge and in moral decisions. Together these result suggest that higher order monitoring appears ubiquitous in various decisions domain resembling real-life choices.

## **2.5. Effects of confidence**

The concept of procedural metacognition defines the monitoring of decision reliability as a mean to adjust decision making. The consideration of value into decision accounts for the consideration of the participants goal and learning processes and therefore to account for this higher order process into a larger and functional view of brain and behaviour. More specifically, in chapters 1 and 2 we present how metacognitive monitoring is a central part of informing behaviour either retrospectively or prospectively and in learning a well. In chapter 4 we present empirical result supporting that confidence levels predict choice consistency over time in various value domain and discuss in chapter 6 how subjective metacognitive profile in value based choice might relate to different behavioural or clinical profiles.

## **3. Thesis synopsis**

The present thesis strives to build a wholistic picture of metacognition by building bridges between its function and computation. Testing the claim that metacognition is a thermostat for coherent behaviour, the question is two-folds: on the output side, what evidence supports its function as regulating the cost-benefit ratio of decisions? On the input side, what evidence supports the fact that metacognition monitors how decisions cohere with subjective values?

Whether it is the mental effort in studding, the risk taken in a gamble, or one's own health that is at stake, we argue that the ability to monitor one's cognition provides agents with the ability to remain coherent with their own goals and values. While most research in metacognition relies on carefully controlled experiments and simplified tasks as coming from psychometrics, the present thesis provides a

framework where value-based choices act as a bridge between these controllable and operationalizable tasks for the laboratory together with the life-like conditions where one relies on her own subjective values. More specifically, stepping back, we present through different angles how value-based decisions come together with the opportunity to apply models that can explain important questions in the field, such as how the computation of confidence levels deviates from normative models and is shaped by different types of tasks. In a nutshell, this thesis provides new conceptual and empirical work that demonstrates the importance of accounting for the values of choices to understand the monitoring role of metacognition as a coherence thermostat.

The thesis narrative builds up from conceptual to empirical work which respectively paints a framework where the function of metacognition is of a coherence thermostat and then attempts to demonstrate that the computation of these monitoring signals relies on subjective value. Chapter starts with the edges of this conceptual framework. On one hand, the concept of coherence is defined by both philosophy and economics as a norm towards which behaviour revolves: the match between one's preferences and one's decisions. On the other hand, cognitive neuroscience studies metacognition mainly for the computation of its input as in perceptual tasks: we present the gap which accounting for subjective value fills in this procedural framework. From this outline, Chapter 2 presents a conceptual landscape for procedural metacognition where different monitoring processes support different tunings of cognition. There we discuss how these various metacognitive monitoring and control apply from infants and non-human animals, to psychiatric disorders, up to assumptions of the legal system on which our societies rely. This chapter thereby links different facets of the study of metacognition to their implications in the real world and highlights at the same time the underlying necessity to account for the subjective value of decisions. In Chapter 3, we ask about the place of value input in the metacognitive computation of monitoring signals. We propose a hierarchical computational model that defines the function of both value and perceptual input in guiding a coherent behaviour. We suggest that metacognition is a powerful tuning mechanism for learning that

provides agents with an adequate ratio of flexibility and resilience to achieve their goals given contextual uncertainty. Together, these three conceptual chapters propose a framework where metacognitive monitoring signals are defined by their functions to ensure decision coherence thanks to their access to subjective value.

The second half of the thesis is concerned with empirical work that more traditionally looks at metacognitive monitoring signals by the other side of the coin: their computation rather more than their function. More specifically, this part aims at demonstrating the role of metacognition as a coherence thermostat by testing the ubiquitous computation of subjective value into confidence reports. Chapter 4 builds up on the previous chapter defining metacognition as a hierarchical thermostat that tracks the coherence between a decision and its decision rule. In this chapter we test the extreme scenario where agents have limited knowledge about the rule they are to use to decide. Building on self-consistency theory of metacognition, we demonstrate that both reflective function of theory of mind and metacognition seem to rely on the same implicit heuristic cues to inform the agents about the reliability of their choices. In Chapter 5, we aim to test the other extreme of the axis where the value of a choice relies on a domain very close to one's own sense of identity: moral value. In this chapter, we design a new paradigm of choice inspired by the literature on volition by asking participants to choose amongst pairs of charities in a similar manner as they are to choose between pairs of snacks. Our results demonstrate that confidence reports sensitive to moral coherence in a similar manner as they are as previously demonstrated to hedonic coherence. While not demonstrating causality, our results also suggest that confidence levels predict the consistency of moral choices over time, supporting the claim that metacognition might act as a coherence thermostat across various types of value domains and tasks. In Chapter 6 we step back from the function of confidence at the decision level to investigate its working at the individual level and ask: if metacognition is an ubiquitous coherence thermostat relying on value (defined by the agent goals and experience), then does it have a subjective criterion? Building on bounded rationality theory, we test whether metacognition appears more as an optimal monitor or a satisfiable one. Our preliminary data suggest that across value domains,

agents appear to present an idiosyncratic sensitivity to coherence as a metacognitive finger print. We discuss the implication of this subjective trait in healthy and clinical population.

To conclude, we present from different angles how metacognition tracks the value of choices in an ubiquitous manner, not to ultimately reach optimal behaviour but instead in a pragmatic manner to inform a thermostat of cost-benefit and remain coherent with one's subjective values. In chapter 7, we discuss the implications of this multi-facet architecture that monitors value as supporting the agent's coherence with herself and with a group.

## References

- Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., & Camerer, C. (2017). Willingness to Pay and Willingness to Accept are Probably Less Correlated Than You Think. *National Bureau of Economic Research*. <https://doi.org/10.3386/w23954>
- Davidson, D. (1982). Rational Animals. *Dialectica*, 36(4), 317–327. <https://doi.org/10.1111/j.1746-8361.1982.tb01546.x>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8(JULY). <https://doi.org/10.3389/fnhum.2014.00443>
- Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1), 5–20. <https://doi.org/10.2307/j.ctvvh84xp.15>
- Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences*, 24(5), 349–362. <https://doi.org/10.1016/j.tics.2020.02.007>
- Hoffmann, S. (2016). Bridging the Gap between Perception and Cognition: An Overview. *Performance Psychology: Perception, Action, Cognition, and Emotion*, 135–149. <https://doi.org/10.1016/B978-0-12-803377-7.00009-0>
- Mill, J. S. (1863). *Utilitarianism* (S. and B. Parker (ed.)).
- Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, 26(C), 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Sepulveda, P., Usher, M., Davies, N., Benson, A., Ortoleva, P., & Martino, B. De. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *BioRxiv*, 2020.04.14.031971. <https://doi.org/10.1101/2020.04.14.031971>
- Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, 18(4), 186–193. <https://doi.org/10.4324/9781315630502>
- Shekhar, M., & Rahnev, D. (2020). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, 1–12. <https://doi.org/10.1016/j.tics.2020.10.007>
- Smith, A. (1986). On the Division of Labour. In P. Cl (Ed.), *The Wealth of Nations, Books I–III* (p. 119).

# Bridge:

## From introduction to landscape.

This thesis argues for the role of subjective value in providing a comprehensive picture of procedural metacognition as a thermostat for decision coherence. By developing the models of both the function and the computation of metacognitive monitoring signals, we suggest that subjective value is essential to close the loop and provide a comprehensive understanding of metacognition. The first half of the thesis presents a conceptual framework for the function of metacognitive monitoring signals.

In this first chapter, we aimed at providing the essential elements in the field of value-based metacognition. As for a puzzle, we started by defining procedural metacognitive monitoring by revealing the edges of the territory: on one side we suggest a function for explicit metacognitive reports and on the other side we define where the research fields stand in this explaining this phenomena. The essence of the gap, we argued, relies on accounting for subjective value: we suggest that accounting for the value of choices draws a holistic picture of metacognitive monitoring by bridging the computation of its signals to their function in ensuring the coherence of cognition and behaviour.

More specifically, we suggest that an agent with explicit access to the coherence of his choice can both ensure the coherence of his decisions but also coordinate them with other agents. We define decision coherence as a concept - the parallel between personal preferences and decisions – and its operationalisation for empirical studies. We also present key concepts of the operationalisation of the research in metacognition, both in perceptual and value-based decisions and highlight the hole in the middle of the picture: how do confidence signals ensure behavioural coherence? We presented key contributions to this question from the field of value-based metacognition and introduced how this thesis provides elements of the remaining gaps.

Resulting from this outline of a map of the known and unknown territories of knowledge in the field of metacognition, we will now attempt in Chapter 2 to better define metacognitive signals not only by their computation as mostly studied in cognitive neuroscience but also by their function as suggested for instance in philosophy. From this multi-disciplinary approach, we will refine a pre-existing concept: the multi-dimensional metacognitive landscape. To do so, we review existing literature to attempt to draw the edges of different elements within this multi-dimensional space and propose a lexicon for these concepts and their relations to each other. Following the two-sided view on procedural metacognition that drives the thesis, the question we attempt to answer here is two folds: what functions do metacognitive signals serve and how does research define their cognitive computation?



## Chapter 2

# The metacognitive landscape, a multi-disciplinary challenge.

### Lexicon:

**metacognition:** (level-two) cognition about (level-one) cognition.

**procedural metacognition:** set of processes monitoring and controlling mental and behavioural states for flexible tuning of performance.

**metacognitive landscape:** proposed theoretical framework to observe various metacognitive functions along three cognitive axes of executive control, representation and conscious access.

**precision:** belief uncertainty that tunes its updating according to Bayes theorem.

**prediction error:** binary expectation about the outcome of a decision as successful or not in reinforcement learning

**decision reliability:** quality of a decision (or set of decisions) as fulfilling a goal. The goal can be bounded rational as requiring a level of success for a level of effort. In a normative sense, reliability can be accounted at two levels: globally as the maximisation of the goal by choice of the decision rule (or strategy, in learning tasks); locally as the maximisation of the decision rule by the choice of the item (i.e. coherence or accuracy in preferential or perceptual task).

**epistemic feeling:** propositional states that suggest to an agent the likelihood of success of her decisions and can guide behavioural adaptation without requiring concepts or awareness.

**confidence:** explicit report about one's subjective appraisal of a decision as correct (i.e. as either accurate or best given a set of options)

**rational behaviour:** behaviour coherently optimizing the utility of one's decisions in regards to her preferences while accounting for uncertainty (e.g. probability, risk, volatility..)

**reasoning:** explicit access to the intention guiding a decision.

**autonomous agent:** agent with the ability to voluntarily and rationally define a self-narrative and to decide coherently a ranking of preferences together with both long and short term plans (i.e. higher-order decisions). The resulting coherent set of beliefs and behaviours enables a fruitful cooperation within a social group (c.f. concept of legal responsibility).

**dual process theory:** framework in psychology dividing cognitive processes into two categories as either unconscious, fast, efficient and relying on heuristic or conscious, slow effortful and reflexive providing the room for deliberate modulation of thoughts or actions enabling to overcome heuristic bias and therefore often result in a more rational behaviour. This definition of decision making systems based on a cost-benefit trade-off is conceptually analogous to our executive control dimension in the metacognitive landscape.

**system 1- system 2 metacognition:** amongst other classifications of the complexity of metacognitive monitoring, the present classification divides them as respectively implicit (e.g. unconscious error correction or epistemic feelings that can be present in non-adult humans) vs explicit (e.g. conscious confidence reports).

**metacognitive thermostat:** concept based on bounded rationality presenting metacognition (in the entirety of its landscape) as a monitoring and controlling system (i.e. procedural) that adjusts a criterion for an amount of resources (e.g. cognitive effort, time, dual process theory...) to be involved in order to reach an intended degree of cognitive or behavioural reliability.

## 1. Introduction

*Truman:* Get in. Look! Shhh...I predict, that in just a moment, we'll see a lady on a red bike, followed by a man with flowers, and a Volkswagen beetle with a dented fender.

*Meryl:* Truman, please....

*Truman:* Look ... Lady ... Flowers! And....

*Meryl:* And... Truman, this is silly!

*Truman:* There it is! There's that dented beetle! Yes! Whoooooooooooooo! Ha-ha! Ha... Don't you wanna know how I did that? I'll tell ya'. They're on a loop. They go around the block. They come back. They go around again. They just go 'round and 'round! Round and round!"

The Truman Show, 1998.

Against what every acquaintance, trusted friend, family member or even his wife assures him, Truman has become certain of only one fact: the world he inhabits is not real, and he must take action to escape it. In the Truman Show, we witness the main character pick up upon irregularities and form a new belief that appears uniquely his by going against what everyone else believe. Incidentally, this new belief reorders his intentions whereby he decides not to continue to live as he always has with the others, but to find, at all cost (so the director ensures), a way to escape this TV set. By dissociating his beliefs, intentions and actions from the popular norm, we follow Truman as he becomes a maverick.

But what ability enables this young adult to question the reliability of his entire world and decide what he ought to believe for himself? What skill enables him to take into account the way the rain falls particularly irregularly one evening but to discard his mother's photo album displaying a rather coherent world? Whatever may be the cognitive process supporting this behaviour, as the audience, we identify ourself in

this hero who decides against all odds to invest all he has in what he is now unshakably certain about.

As humans, our metacognition enables us not only to appraise the validity of simple decisions, but to take into account the probabilistic reliabilities of facts and hence reason rationally upon our own beliefs and intentions. If this ability has been singled since Socrates, to this day, this hallmark of human rationality is still being defined and investigated at large throughout different fields.

In this review, we summarise the literature in the fields of philosophy and psychology of metacognition and highlight how both field's main approaches contribute to complementary understanding of this reflective function. Here we focus on procedural metacognition which monitoring aims at adjusting cognition or behaviour. While both field cover a wide arrays of topics of research under the umbrella term of "metacognition", we here thrive to present how across both fields these different and often barely related perspective appear to come together on a simple idea that metacognition appears as a common currency of cognitive and behavioural reliability. In a nutshell, we try to argue across this wide research that all comes together under the image of metacognition acting as a thermostat and adjusting the agent's use of resources in the aim of being successful. Building on a simple proposal of three dimensional metacognitive landscape a decade ago, we come back to a "metacognitive landscape" (Fig. 1, Fleming et al., 2012) and discuss in both fields the concepts that shape it.

First, we present how philosophy defines metacognition along the tree axes of this landscape by situating it in regards to meta-representation, conscious access and executive control. In line with our restriction to procedural metacognition, we define its role in regards to the most popular theories about the function of the brain: learning. We furthermore discuss the links between metacognition and the social imperatives of the human condition and extend on its role in complex and social tasks which we encounter in real world.

In a second part, we present how the empirical research in cognitive neuroscience studies metacognition in regard to this metacognitive landscape. Focussing on metacognition of decision making, we start by presenting the contemporary and popular normative model of confidence. Presenting the cues on which this reflective

mechanism relies, we then open onto other models and functions of metacognition that, by focussing on different task, suggest different monitoring for different executive controls. Most of all, we present how, besides retrospective error correction, metacognition appears to be embedded in a prospective role for adjusting resources allocation to ensure success. In line with this thermostat view of procedural metacognition, we highlight the multi facet role of metacognition in a bounded-rational framework whereby learning about the global task goes hand in hand with metacognition. Finally we summarise some literature on the evolution, development and disorders of metacognition in view of our metacognitive landscape. Together this overview of the field of metacognition in both Philosophy and Cognitive neuroscience portrays its role as a regulator of cognitive and behavioural reliability from simple decision-making to shaping an autonomous and coherent agent within a society. We highlight the gaps in the literature based on this landscape and the different areas of research that constitutes it.

## **2. Philosophy of metacognition**

What would it change to make robots metacognitive? Would they be smarter, conscious, better at learning or at collaborating? To really grasp the nature of the metacognitive function, philosophy thrives at defining the core mental functions it brings to the table. Building on the common conception that the brain is a predictive machine, what does a recursive mental function bring to an agent? What is its epistemic role and value? While the literature on philosophy of metacognition is large and sometimes empirically informed, we present here how all these various areas of research appear to come down to a few key dimensions: a “metacognitive landscape”.

### **2.1. A meta level landscape**

#### **2.1.1. Learning**

The brain is a central network of neurons towards which most of the animal’s perceived information converges, and from where most actions are guided. A widely acknowledged theory sees this organ as a predictive machine which accumulates evidence and learns states and causal structures of the world to adapt its actions to

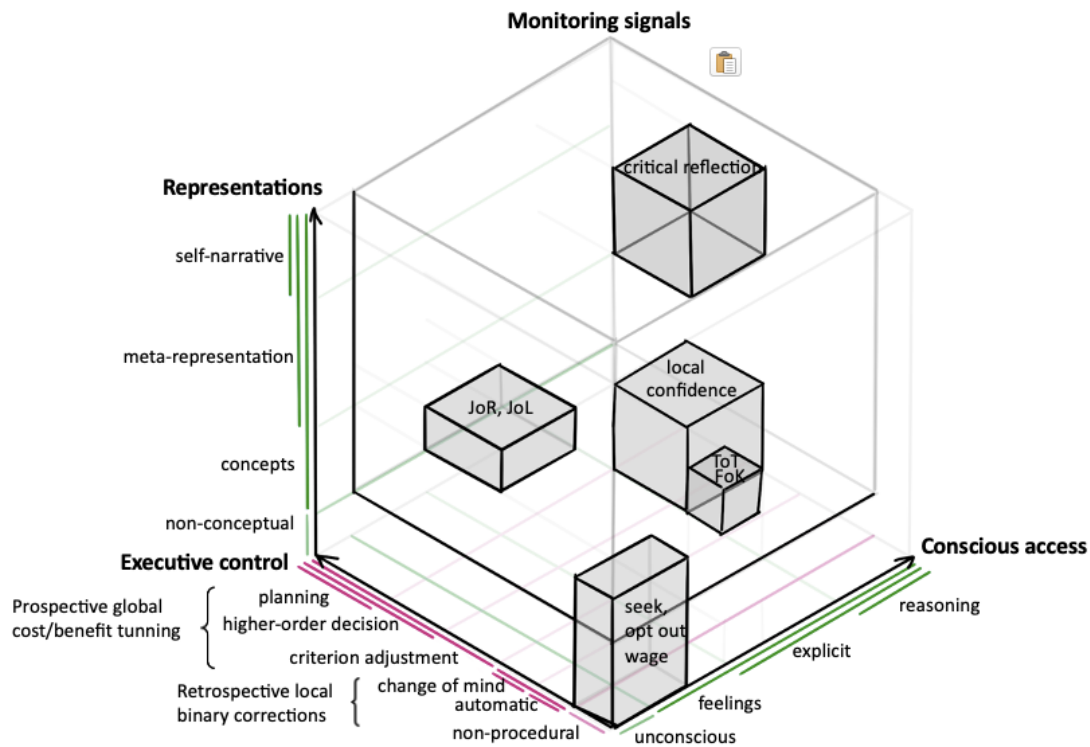
it (Hohwy, 2013). But if the central function of the brain is to learn about its environment, what does metacognition have to do with it? Is it always present? Is it advantageous?

The most popular framework to describe its working is Bayesian inference. In a simplified formula, predictive processing suggests that a learned prior expectation is updated by incoming information either in the form of sampled evidence or feedback from one's action (Clark, 2016). This simple model predicts that learning or an update only occurs when the model at hand is at odds with the incoming evidence. This difference or error is then corrected within the model. Beyond this binary correct – error updating, the Bayesian model does account for uncertainty. By weighting both the reliability of the expectation and of the incoming evidence, the update can be advantageously refined. This reliability in the predictive processing literature is known as precision (Hohwy, 2012). Since both the evidence of the expectation and incoming evidence together with their uncertainty can be read out and cognitively to update the model, such Bayesian learning can be seen as “level one”. Such a learning model has also been demonstrated in non-human animals (Lak et al., 2017).

### *Second order monitoring*

In Philosophy more so than in empirical science, the debate is at where to draw the line between a cognitive and metacognitive function. The definition seems simple: a cognitive or level-one processing takes incoming evidence from the world whereas a metacognitive or level-two processing is recursive in the sense that its incoming evidence is the cognitive process itself or its output. This “aboutness” of the second order is referred to in the philosophical literature as attributive, of a cognitive process or content. As varied as the cognitive processes themselves, defining what makes a process metacognitive can be challenging and take various dimensions. A useful concept is therefore to talk about metacognitive landscape such as expanding along various dimensions. A simplified model of metacognition was proposed to extend along 3 axes, namely meta-behaviour as the monitoring and control of behaviour as mainly studied in cognitive neurosciences ; meta representation as a

conceptual representation of a first order content ; and consciousness as for the level to which one will be aware of this second order output (Fleming et al., 2012). Focusing on the implications of a second order here and its implications for behavioural optimisation, the other two dimensions will be discussed in the next sub parts.



**Figure 1: a metacognitive landscape** (adapted from Fleming et al., 2012) Metacognitive research can be seen as defining its workings and function along the 3 dimensions that constitutes it. The executive control (pink) of cognition and behaviour or “decisions about decisions” mainly studied in cognitive neuroscience distinguishes procedural metacognition (e.g. Judgment of Learning (JoL), Judgment of Rightness (JoR)) from non-procedural (i.e. content) metacognition (e.g. Feeling of Knowing (FoK), Tip of Tongue (ToT), Proust, 2010). Here we suggest to distinguish two control function as retrospective for correction of past decisions or prospective based on a cost-benefit ratio that relies contextual knowledge. Increasing with complexity of these control functions, metacognitive monitoring require increasing complexity along two other axes: representational (blue) and conscious (green) access. The complexity of these functions of procedural metacognition increases in evolution and development by intertwining the three different axes. The efforts of

both philosophy and cognitive neuroscience contribute to understanding the links and dissociations of the different components of this metacognitive landscape.

### *Free energy principle (FEP)*

As another theory about the brain, free energy principle suggests that the brain emerged as other thermodynamic systems to regulate some imbalance within the individual homeostatic states but also eventually also with the environment (K. Friston, 2011). This theory aligns with the predictive mind theory described above but furthermore echoes distinctively with the notion of confidence monitoring. Indeed, beyond learning to optimise one's behaviour to a given context, according to FEP if the brain emerges to minimise the incoherence within the agent's states and beliefs and with its context: measuring this incoherence or free energy appears as a powerful catalyser to reach this end (Henriksen, 2020; Moulin & Souchay, 2015). Pushing this theory further, it could be argued that by monitoring such incoherence itself, this higher order is a key catalyst to provide an animal with the autonomy to regulate itself and its fit within its environment. FEP defines distinct types of monitored FE such as epistemic and motivational uncertainties that respectively refer to reliability of beliefs about the states of the world and about its causal contingencies (K. Friston et al., 2015; K. J. Friston, 2018; Pezzulo et al., 2018). By defining further the operationalisation of cognitive types of free energy, computational simulations or cognitive studies could be applied to demonstrate the role of such higher order monitoring in the adaptive or autonomy of an agent in a simple context (Yon, 2020).

### *Theories of metacognitive rational behaviour*

Another way of looking at this ability to control the coherence of one's thoughts and actions in a self-coherent manner is seen as rationality. In broad terms, rationality can be defined as behaviour maximising one's utility (*i.e.* defined by one's goal) while accounting for costs and risks (*e.g.* volatility, probabilistic uncertainty...). But what are the mental capacities necessary to display such rational behaviour, and is metacognition part of them? A central aspect of the debate relies on accounting for one's uncertainty. The philosopher Davidson suggests that, beyond the ability to

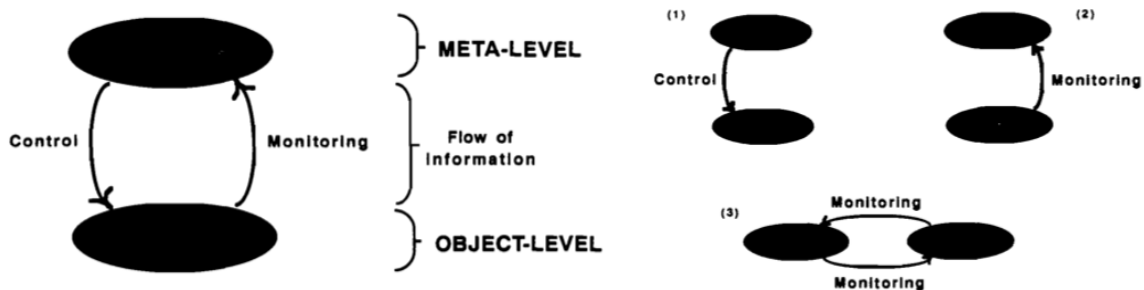


verbally report our own mental states, the ability to evaluate our mental states (*i.e.* to have thought about thoughts) is key to human rationality. This debate is currently fed by ongoing research which demonstrates that infants and non-human animals behold the ability to access their own uncertainty about their actions. By opting out of difficult decisions, non-human animals have arguably demonstrated abilities to coherently align their decisions to the reliability of their beliefs to maximize rewards (*i.e.* rational behaviour). However, the need for a second order monitoring or even conscious access by these animals to skilfully deal with uncertain stimuli has been greatly controversial for its requirement of a second order monitoring. To overcome the key limitation of these study where the response to uncertainty was linked to the amount of reward associated to success, Smith et al. (2006) dissociated the cues for reinforcement learning and reward from the cues associated with uncertainty and demonstrated that one of their two monkey appeared to monitor uncertainty of his responses without the need for incentivised motivation. More of the empirical findings demonstrating the role of metacognitive monitoring in behavioural control will be discussed in part 3.2, and its relation to other aspects of rationality such as reasoning in part 2.1.3-4.

### **2.1.2. Meta representation**

Besides the role of metacognition in learning and behavioural control previously discussed, one can ask: what are the required mental capacities for metacognitive monitoring? Do non-human animals behold them? Do some humans lack these abilities for self-monitoring? To answer these questions, Joelle Proust distinguishes two metacognitive concepts for their different function. First, declarative metacognitive knowledge is the descriptive knowledge about one's own cognitive processes, their workings and factors that influence them (e.g. knowledge about one's learning own methods). Second, procedural metacognition which comprises metacognitive monitoring as the evaluation of one's own ongoing cognitive activity (e.g. "Did I learn this material enough?", "How likely am I to succeed?", "Should I correct this decision?") and monitoring control regulating this ongoing cognitive activity (e.g. selecting the study material, allocating effort or terminating a task, correcting a task) to result in improved mental or behavioural performance (Kluwe,

1982; Proust, 2010; Roebbers, 2017). This functional difference dissociates metacognitive monitoring from metarepresentations: while a meta representations can be a passive read out of a first order process, the pragmatic end of a metacognitive process requires an engaged query that resamples evidence in a normative manner by assessing the adequacy of the cognitive process. As a key result, this distinction of metacognitive monitoring from metarepresentations liberates it from the latter’s conceptual requirements for concepts, language or even the ability to represent one’s or others’ minds. This pragmatic definition of metacognition can then accounts for seemingly metacognitive behaviours observed across the animal kingdom and in many instances of daily life such as error correction which corrects behaviour while being implicit and not meta representational. Interestingly, one of the early definition of metacognition also put the emphasis on this procedural definition by arguing that a second order cognitive system was defined by an ability to top-down control the lower cognitive system, and distinguished this higher order cognition from higher order representation which can be seen instead as parallel systems (Fig 2).



**Figure 2: procedural metacognition, a meta level defined by monitoring and control** (taken from (Thomas O. Nelson & Narens, 1990)). The left panel represent the defined meta level as supervising the object level cognition. The right panel presents the complementarity of both monitoring and control functions, with the possibility of independent system: 1) controlling without monitoring; 2) monitoring without any control or even 3) the lack of necessity for a meta level if top down control is not present and the possible parallel monitoring of different cognitive systems. While the early studies of metacognition in metamemory investigated the role of monitoring in behavioural control, the more recent perceptual studies focus

more importantly on the monitoring side while the study of behavioural control is often seen as mainly investigated by the executive control literature (Roebbers, 2017).

### *Epistemic feelings*

Can metacognitive monitoring cue an agent without requiring a meta representation? What creatures could have it and are the required mental capacities for it? Noetic or epistemic feelings are propositional states that suggest to an agent the likelihood of success of her decisions and can guide behavioural adaptation without requiring the conceptual awareness of one's success or failure. While feelings are considered to rise above the threshold of consciousness, the requirement for a normative notion of goal or success can be approximated to heuristic appraisals such as feeling of ease, or familiarity. The literature on metamemory was essentially built on epistemic feelings such as Feeling of Knowing (FoK: could recognise if given multiple alternatives) or Tip of Tongue (ToT: about to retrieve the full information).

### *Meta representations and metacognition*

What do representations at the metacognitive level bring to these appraisal? Representations, or conceptual thinking can be seen as mental objects which are useful for building and organising precise mental maps to guide behaviour or share with a social group. Therefore, unlike epistemic feelings based on heuristics, evaluative meta representations can come as appraisal of one's response accuracy or likelihood to succeed at a given goal. As a pragmatic tool, such conceptual representations of reliabilities can be seen as useful mental objects available to the conscious level to trade between measures of reliability and guide behaviour. Such strategic use of explicit representations in guiding behaviour will be more extensively discussed in part 3.2, but can as an example help to steer one's behaviour towards a task where one is more likely to succeed (De Gardelle et al., 2016; Lee et al., 2021; Rouault et al., 2019) or be communicated to boost the group decision toward the answer that one sees as most reliable (Bahrami et al., 2012). This dimensionality of meta-representation for metacognition opens two questions. First, how this complexity evolves in infants as whether it requires language,

conceptual thinking or theory of mind is up for debate and will be extensively discussed in part 3.4. Secondly, how do these propositional feelings inform our conceptual and explicit levels of confidence in our decisions? We discuss how physiological cues related to self-monitoring can contribute to explicit metacognitive reports in the following part.

### **2.1.3. Consciousness**

#### *Content metacognition for consciousness*

As discussed above, the early days of metacognitive research in metamemory was heavily anchored in epistemic learning and referring mainly to the question of monitoring the presence of a cognitive content. This link between metacognition and meta level or higher order representation about first order processes led to the suggestion of a need for metacognitive read out and access to become conscious of one's state. Higher Order Thought theory suggests that one must be aware of being in a given state to be conscious of this state (Rosenthal, 2005). The use of metacognitive decisions such as wagering were also used as cues to distinguish conscious decisions from guesses (Seth et al., 2008). However, as for the distinction made by Proust between procedural metacognition and meta representation (Proust, 2010), so can it be done with procedural metacognition and consciousness (Fleming et al., 2012). Nelson responds to Rosenthal's proposal by highlighting: "is there a difference between one's state being conscious of another state vs. one's state "monitoring" another state?" (T. O. Nelson, 2000). To answer this question, we dive into the technical definitions of consciousness and metacognition to this day.

#### *Procedural functions of consciousness and metacognition*

In Dual System Theory, Kahneman (2003) presented a dichotomy of decision processes: the first system makes fast and efficient decisions, and, while the incoming percept and output choices can be consciously accessed, the decision process itself based on heuristic is to the participant as a black box: inaccessible. The second system on the opposite makes – when favourable - slow and more costly decisions by overcoming the instinctive heuristics and result in a more rational behaviour. In this system, the decision process relies on reflection and is therefore

transparent to the agent who can explicitly report about the factors of the choice. According to this dichotomy, decisions are therefore both reflective and conscious, or on the opposite reflexive and unconscious. On the metacognitive side, the reflective nature of system-two was proposed to rely on the metacognitive monitoring of the system-one output to be overwritten necessary by the most costly decision system (Thompson, 2009). On the consciousness side, the Global Workspace theory (GWS, Dehaene et al., 1998) suggests that consciousness emerges from the broadcasting of evidence into a cognitive space from multiple brain areas to enable an integrative singular experience of the world. This theory proposes a kind of cognitive blackboard where evidence and thoughts can be rearranged and manipulated aligns both with Kahneman functional definition of consciousness and cognitive concepts such as working memory. In our metacognitive landscape, we propose this multi facet function of metacognition as being able to both monitor the accuracy of local decisions and also to prospectively adjust the decision making system to ensure its reliability. In line with this procedural reliability, we propose that metacognition acts as a thermostat which tunes the amount of resources to be invested in a choice to ensure it meets one's goals. In other words, procedural metacognition as defined before aims at monitoring the reliability of cognitive processes (rather than their content) in order to adjust them and maximize success at the task at hand. We know nonetheless that metacognitive monitoring and consciousness can come apart (as orthogonal dimensions) such as in dreams where one is conscious but without reflective ability or when one experiences an illusion while knowing that this percept is inaccurate. Below, we discuss what consciousness can bring to metacognition and *vice versa*.

### *Procedural metacognition levels and consciousness*

As previously discussed for the axes of behavioural control and meta-representation, the metacognitive processes (both monitoring and controlling) can be defined for incremental levels of complexity. Here we discuss some categorisations of metacognitive processes and their relations to conscious access. As a theory unified with free energy and active inference, Timmermans (2012) proposed that the human brain models 3 loops of interactions: an internal or inner loop that thrives at

representing itself and predicting how actions in one brain region will affect other brain regions; a perception-action loop popular in active inference whereby the brain tried to model the consequences of certain actions depending on the perceived states of the world; and lastly an outward or social loop that predicts the reactions of others to our own thoughts or behaviours. The authors suggest that, as an independent dimension, conscious access can improve the monitoring and control of all 3 loops. Interestingly, J. Metcalfe (2012) defines 3 types of metacognitive judgments as inspired from 3 types of conscious access: anoetic, concerning objects in the world; noetic, concerning mental representations; and auto-noetic, in which the referent includes the self. It can be noted that although coming from different theories, namely FEP and consciousness, both these hierarchies of metacognitive judgments present some interesting functional overlaps while suggesting different relations to consciousness. One last popular classification of metacognitive judgments refers to Shea social evolution theory of metacognition (Shea et al., 2014) which labels implicit (non-verbal) metacognitive monitoring as “system 1” and explicit monitoring available for verbal report as “system 2”. This role of social interaction in shaping explicit reports of metacognitive judgments such as confidence levels will be discussed in length in the next part.

These theories about classifications of metacognitive monitoring are essential to our understanding of its evolution, development and eventual training as will be discussed in part 2. Nonetheless, the various subfields of cognitive neuroscience need to provide dedicated research to test empirically the evidence in support of such computational models for metacognition. Aligning with Timmermans classification describing conscious and unconscious metacognitive monitoring as spanning from an inner to an outer loop, it appears essential to test for instance whether interoceptive cues could work as metacognitive signals monitor and adjust physiological states. The work of Micah Allen has been particularly devoted to this endeavour by investigating the links between interoception and metacognition together with a unifying theory of the brain. Within the framework of embodied inference, empirical evidence suggests that by representing the expected outcome of a decision, emotional valence can work as a vector to optimise confidence reports

(Hesp et al., 2021). By furthermore demonstrating the effect of heart rate or arousal on the buildings of confidence levels, Allen et al. suggest a strong contribution of interoceptive signals to metacognitive monitoring. While discussing the altered physiological signals and interoceptive abilities in psychopathologies such as addiction and depression, the authors highlight the possible importance of accounting for these embodied cues in theories and models of metacognition (Allen et al., 2016; Legrand et al., 2021).

Relating to the classification proposed by Shea (2014), empirical evidence also supports that different computational models are at play for different levels of conscious access to metacognitive monitoring. As a starting point, it was demonstrated that humans can monitor and adjust their behaviour unconsciously in daily tasks such as keyboard typing (Logan & Crump, 2010). In neurobiology, the cerebellum is also famously notorious for its elegant neuronal architecture, which, as a distinct brain region from the rest of the brain, enables it to quickly and autonomously correct motor actions. But while as notorious multi-taskers we are all aware of our ability to perform quite complex tasks in “autopilot mode” and without the need for conscious awareness, when and why do we consciously engage in the conscious monitoring and adjustments of our thinking or doing? As discussed above, we broadly defined consciousness as providing the ability to think or act deliberately. As will be discussed in part 3.2, consciousness can access independently level-one cognitive processes from level-two monitoring processes. An elegant example of the implication of conscious monitoring (or insight) can be taken from the clinical literature. Anosognosia describes a state in which patients are unaware of their illnesses and it has dramatic implications on patients ability to deal with or recover from their condition. It goes without saying that the demanding task of adjusting one’s beliefs and lifestyle to live with a clinical condition requires one to be conscious of these limitations. Katerina Fotopoulos dedicates her work to understanding insight in clinical patients and demonstrated that by using the help of videotaping, she could restore conscious awareness of an anosognosia patient who could otherwise not predict her inability to move her paralysed arm (Fotopoulou et al., 2009). While this example is quite drastic, it illustrates clearly the

implications of conscious metacognitive monitoring on the possibility to deliberately adjust one's thoughts and decisions for daily life. But what are the key cognitive processes supporting this conscious metacognition? If we approximate conscious metacognitive monitoring to metacognitive monitoring available for verbal report, then one could suggest that metacognitive feelings that make it to the conscious threshold would be available for report. To relate it to a more functional picture of consciousness, we can go back to the GWS framework which suggests that this conscious space both enables broadcast of evidence across domain-specific cognitive systems (perception, value, memory..) and make these monitoring available for verbal reports. The question then remains: what is necessary to pass this threshold of consciousness? And how does this conscious access in turns then profits to the procedural function of metacognition?

#### *Procedural metacognition for consciousness*

What makes information consciously aware? The link between confidence monitoring and attention has extensively been demonstrated where low levels of confidence boosts the recruitment of attention (more extensively discussed in empirical part 3.2 on metacognitive control). A recent theory by SM Fleming also suggests that the determinants of whether the information is relevant enough to reach the consciousness threshold in part rely on metacognition: besides the recent demonstration that perceptual attention is guided by goal directed sampling process, the Higher Order State Space (HOSS) theory suggests that the absence of an object can become consciously aware due to the difference between the certainty of an expectation to find an item and the certainty in the sampled evidence. In other words, consciousness is a top down process weighted by metacognitive certainty of expectations. Aligning with this theory is the Metacognitive Reasoning Theory which suggests that for a task such as in a Cognitive Reflexion Test (CRT) to involve conscious reasoning rather than heuristic, an initial heuristic metacognitive monitoring evaluates the need for recruitment of the effortful and slow conscious reasoning to solve the task at hand. According to these views, we could suggest that consciousness could therefore (at least in some cases) be regulated by a sort of metacognitive thermostat that would monitor the reliability of an automatic



response for the present task and, if need be, recruit costly consciousness to provide a more reliable and rational conscious response (or belief formation) than the fast initial one. The links between metacognition and mindfulness will be discussed with the empirical literature for metacognitive control in part 3.2.

Beyond this suggestion that reaching concisions state relies on reliability monitoring, Shea recently suggested that the nature of consciousness as a global workspace requires that all evidence passing this threshold should be weighted by a reliability tag as provided by metacognitive monitoring . Indeed, according to this view, information that is integrated in a singular conscious experience requires a common scale of reliability. Overall, we therefore discussed that the interaction between metacognition and consciousness seems for the least bi-directional and involving different levels of complexity such as defined by their interaction along the axes of metacognitive landscape. While the complexity of metacognitive monitoring as defined in the landscape is dictated by its procedural function (*e.g.* automatic motor adjustment, change of mind, long term planning), in the following part we discuss how such complex behaviour appears intertwined with the social origins of human evolution.

## **2.2. Social origins**

We discussed the theoretical classifications and dimensions of a metacognitive landscape. In this section, we focus on the evolutive theories that thrive to argue for the emergence of metacognitive monitoring as conscious and available for explicit report. We discussed Joelle Proust's argument for a distinction between attributive meta representation and evaluative metacognitive monitoring (*i.e.* procedural). Here, we first touch upon the theoretical suggestions about the roles of this explicit monitoring for social functions before discussing its associated mental abilities.

### *Social function of explicit metacognition.*

Social learning suggests that while evolving in social groups, humans developed the ability to learn from other agents by identifying the reliability of their behaviour or views. An experiment demonstrated the heuristic association of others' slow response with a low reliability of their response (Patel et al., 2012). Language has

unquestionably evolved as an advantageous vector of information exchange. In parallel to this evolution, it was suggested that the metacognitive ability to communicate levels of confidence might similarly result from cultural evolution (Heyes et al., 2020). In other words, explicit levels of confidence can therefore be seen as a between-agents currency of evidence reliability, similar in nature to the previously discussed Bayes-like between-cognitive domains computation of the GWS. Communicating levels of confidence in one's beliefs or behaviour was indeed demonstrated to often contribute to improved quality of group decisions and collaboration (Bahrami et al., 2010, 2012; Hertz et al., 2017). Furthermore, the demonstration that one's confidence levels recalibrate itself depending on the social environment to improve group decision is strong evidence in support of this hypothesis (Hertz et al, 2017).

Another parallel suggestion of the role of explicit self-monitoring is that by being aware of one's beliefs, desires and actions, one can build a coherent narrative and behaviour such as by efficiently planning, executing, evaluating or cooperating. Such intellectual and practical insight is argued to be a central building block for rational and social agency by providing an individual with the inner coherence necessary for fruitful coordination and social interaction (Greco, 2019). We previously used the analogy of metacognition acting as a thermostat to provide the agent with conscious awareness in order to ensure reliable and coherent beliefs and behaviours. Building on this analogy, it could therefore be argued that the resulting self-awareness could have been cultivated over generations by tasks requiring a high level of autonomy and foresight as when becoming part of a functional and thriving social group. But whether metacognitive insight provided agents with the ability to collaborate fruitfully or whether these social interactions did tune these self-monitoring and controlling abilities is still to this day unanswered. One thing for sure is that in this fast-paced interconnected world with ever growing quantity and complexity of opportunities and social interactions challenges our metacognitive abilities (see part 3.3 for metacognitive training and extensions).

*Metacognition and theory of mind*

To best understand the workings of metacognition as providing self-awareness, theories about its origin thrive to identify the tasks and mental capacities with which it can be associated. As a converging hypothesis, explicit metacognition and Theory of Mind (ToM), or the ability to dissociate one's own mental states from those of others, are often suggested to share common origins (Frith, 2012). Timmermans goes in depth to suggest that in order to know how to deal with other minds efficiently, we also need to know how to deal with our own (Timmermans et al., 2012). In his argument, the author suggests that self-awareness should naturally emerge together with social skills as one of 3 inference loops (see part 2.1.3 procedural metacognition levels): as an interaction between the abilities to monitor and regulate one's own states altogether (inner loop) and together with the outside world (perception-action loop), the ability to monitor one's mind in relation to others' minds could have emerge as a functional set of skills. But what would the implications of having both functions being interdependent? Does that mean that animals who are not social or able to differentiate their mind from others cannot have self-awareness? And if an agent loses the ability of self-awareness does that make her lose the ability to conceive that others have minds of their own? In other words, what are the evolutionary, developmental and clinical implications of such suggestions? By dissociating the philosophical notions of self attributivism from the notion of self-monitoring, Joelle Proust provided a solid philosophical distinction between both concepts (Proust, 2010). However, in empirical research, the similarities and distinctions between both mental capacities are still generating great debate. The empirical literature will respectively be discussed in the following parts about developmental (part 3.4), clinical (part 3.5) and animal studies (part 4.2) of metacognition.

### **2.3. Responsibility**

#### *Sense of agency*

As for metacognitive monitoring, sense of agency provides agents with a subjective evaluation of their causality in the outcome of their action. Bigenwald and Chambon (2019) stress this subjective experience by differentiating the concepts of “having a choice” and “making a choice”: the first is to have the objective options available, the second is to have the subjective disposition and psychological ability to choose amongst these options. Sense of agency comes with a feeling of control over one’s behaviour and outcomes. The notion of agency mainly arises from the philosophy of free will where the debate aims at identifying whether a sense of agency reflects a real psychological ability to voluntarily decide or whether our biological brain determines our behaviour and the sense of agency only provides us with an advantageous illusion of selfhood to protect and integrate within a social group. The famous “choice blindness” paradigm however seems to suggest that sense of agency could be a construct to prevent cognitive dissonance: when presented with 2 alternatives, we might be very good at justifying why we believe we chose an option even if this latter was changed for the options we rejected without us knowing it (Hall et al., 2010; Nisbett et al., 1977). Even if we do not manage to track our own choices (choice blindness), we seem skilled at creating a narrative to justify how coherent these choices are with our self and beliefs, therefore rationalising our self-evaluation. This research opens questions on the link between two facets of choice evaluation: How does our ability to track the choices we make inform the evaluation of our choices? And does more self-knowledge in turn improve an agent’s ability to track the choices she makes (*i.e.* reduce choice blindness)?

#### *Self-knowledge and self-control*

In his study of free will, Frankfurt defines a hierarchy of desires: first order desires are about objects in the world such as wanting to eat the marshmallow, and second (or higher) order desires which are desires about first order desires such as wanting to patiently wait without eating the marshmallow. With this concept, Frankfurt argues that conflicts between orders of desires can make us “choose what we do not want and want what we do not choose”. This distinction according to Frankfurt

defines different levels of self-control and can account for certain impulsive or addictive behaviours (Frankfurt, 1971). While for instance Ulysses famously tied himself to the mast of his ship to prevent himself from being attracted by the sirens' songs, to our day many alternatives can help us stay on track with our higher-order desires such as by restricting app usage, or placing a financial bets online for or against our own behaviours. Building on our metacognitive thermostat analogy where consciousness and reasoning can be recruited to make a challenging decision, on a longer timescale, this analogy can be extended as by monitoring the reliability of one's higher-order desires and accordingly recruiting strategies to nudge oneself to align our choices with our actual higher-order preferences. On different scales, metacognitive monitoring can therefore be suggested as a trigger to recruit the mental capacities for the agent to fairly consider all her options in light of her higher-order preferences and beyond first order impulses, giving her a better chance to "make a choice". More empirical research on the role of metacognitive insight in regulating impulsive behaviour or choosing according to one's higher-order preferences could bring light into this complex metacognitive function.

### *Legal justice and self-awareness*

Legal justice defines responsibility by the concepts of actus reus and mens rea: "To be criminally liable, one must (1) consciously will to x; (2) know that x is wrong; and (3) do x" (Bigenwald & Chambon, 2019). Both the ability to be aware of one's higher order desires and to evaluate one's actions are therefore required to be considered as a rational agent who is responsible for her own doings. Once one is found to have committed a crime, the court jury can assess how guilty the agent was based on his mens rea. In this exercise, it was demonstrated that members of a jury often rely on the expression of remorse by the criminal. Bennett (2016) defines "Remorse concerns the person's assessment of their own performance, achievements or standing in some direction or other" and goes to discuss its role in learning from one's experience such as with the illustration by the novel *Crime and Punishment* (Dostoyevsky, 1866) where the main character Rodion, although not imprisoned for his double murder, ends up losing his mind to moral disgust and paranoia due to his

wrong doings. In other words, the feeling of remorse is commonly assumed to act as a window into someone's true higher order desires and intentions, and is therefore taken into account when incriminating an individual as potential for self-correction.

We have therefore discussed that the law sees as responsible people with the abilities to act according to their morals or at least indulge into feelings or wrongdoing otherwise. But if these criteria for self-evaluation and self-control are really central to the legal concept of responsibilities, are there people or circumstances in which these mental abilities cannot apply that exempts one from responsibility? Indeed the law accounts for non-adult individuals in which this metacognitive-like awareness is not considered mature enough to be responsible for criminal acts (Carroll, 2016). Psychopathologies such as schizophrenia where one's ability to distinguish reality from hallucinations are also exempt from legal responsibility (Bo et al., 2015; Van Der Plas et al., 2019), and with an aging population where autonomy or dementia can be a daily challenge, the question of responsibility also arises (Fleming, 2021). On the contrary, our highly complex social construct also attributes higher levels of responsibility to individuals with specific social status (political, military, medical, entrepreneurial, familial...) independently from the individuals' mental abilities for actual responsibility as defined by the law.

Besides these subjective limits in mental abilities, could there be some situations where one could find her self-awareness and self-control challenged to alleviate responsibility? Research in neuroscience and economics could inform the law to best evaluate subjective abilities to "make a choice" in a given situation (Bigenwald & Chambon, 2019). Several studies tend to demonstrate the limits of mental capacities to act rationally and according to one's higher-order preferences in complex environments. Two of those instances can be the reduced sense of agency (ability to feel in control of one's action) when part of a large social group (Beyer et al., 2018), or the discounting of moral values in large and complex markets (Falk & Szech, 2013). Both the ability to remain aware and truthful to one's higher-order preferences therefore appears challenged in complex social environments which we face in everyday life. The study of metacognitive insight in such moral decisions

could provide a better understanding of the working of human rationality and responsibility in the real world. Furthermore, while in ethics, the debate of moral enhancement discusses biological enhancers for moral behaviour in people or challenging situations, scientific studies on the psychological working of metacognition in higher order preference and complex environments could provide a powerful cognitive lever for rational and moral behaviour in our fast evolving world.

### *Self-regulating social groups*

Promoting such reliability monitoring within a social group can optimise its efficiency. While we previously discussed how genuine self-monitoring improved coherence within an individual and collaboration between agents, some social organisations actively foster these self-regulatory mechanisms within their organisation to further boost their efficiency. In science for instance, the replication crisis for instance raising a red flag on the reliability of overall scientific method resulted to the creation of an active self-regulatory procedure with the popularisation of sharing early hypotheses and data across laboratories (Woelfle et al., 2011). Similarly, in the context of law, the reliability of witnesses reports is known to often be biased and the emergence of the field of Neurolaw aims at finding ways to optimise the reliability of such collaborative efforts to accurately incriminate criminals (Bigenwald & Chambon, 2019). By developing self-monitoring abilities of pupils at school, we will also later discuss how such metacognitive enhancement can favour individuals and their societies at large (c.f. part 3.5). Beyond the agents genuine metacognitive self-monitoring, efforts to create systems in which groups can self-monitor appears as an advantageous endeavour.

## **2.4. Non-human intelligence**

### *Who self-monitors?*

To disentangle the multi-dimensional metacognitive landscape as an organised network of higher-order cognitive functions, empirical studies test its function in different tasks and individuals (e.g. age (c.f. part 3.3.2) and clinical spectrum (c.f. part 3.3.3), but also across species (c.f. part 3.3.1). In parallel of the actual findings on the workings of metacognition, the question of its ethical implication emerges. Indeed, the early concepts of metacognition relying heavily on epistemic feelings (Fok, ToT...) seriously questioned the independence of uncertainty and self-monitoring. The question of whether animals like rats who could opt out of uncertain situations therefore opened the question of whether such species also had anything close to self-awareness. In parallel, the question of whether uncertainty monitoring and consciousness are linked together raises tremendously heavy ethical question. Indeed, animal rights heavily reside on the assumption that non-human species do not have consciousness or a concept of self which could provide them with human like physical and mental suffering. The ongoing research on the emergence of higher order cognitive processes such as metacognition in non-human animals therefore bears important ethical implications on our relationships to other living species. As previously discussed in the section on metacognition and legal responsibility, the ability to self-monitor in other species could provide them with tailored consideration and rights.

### *Who should self-monitor?*

Providing self-monitoring and metacognitive traits to artificial agents could be advantageous for their interaction amongst themselves and with humans. Indeed for instance, self-driving cars communicating an uncertain obstacle to their driver or to each other could, as humans, slow down and recruit the necessary resources to achieve a reliable - or here safe - outcome. Nonetheless, in parallel to the previous discussion of the entanglement of non-human metacognition and rights, providing metacognitive traits to artificial agents could bear similar ethical implications. Together with uncovering the link between metacognitive and other higher-order

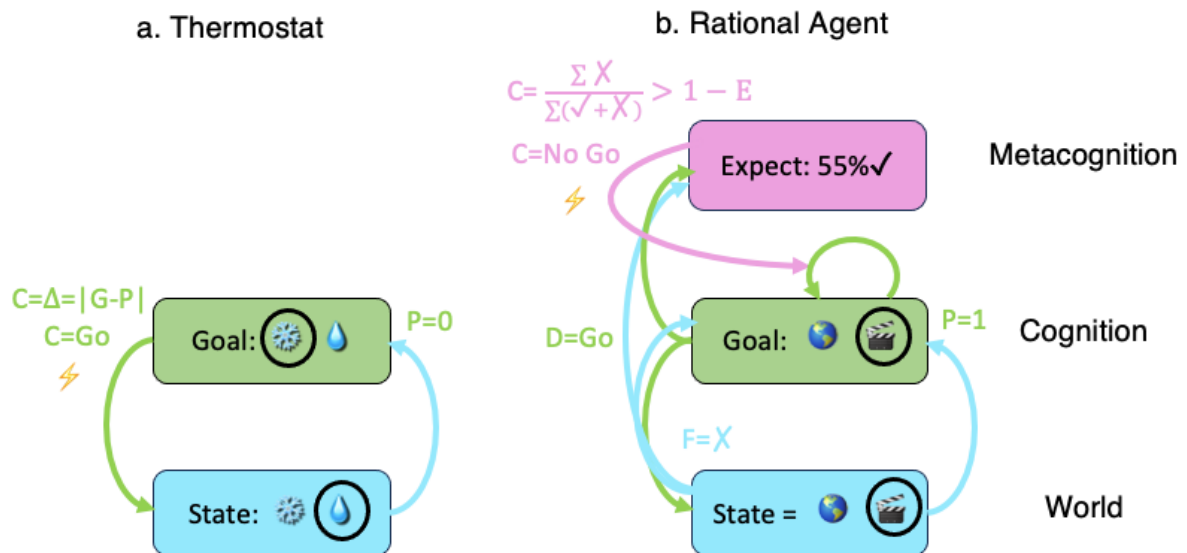


cognitive functions in intelligent agents, the question of etherical consideration for these near human features therefore emerges.

## 2.5. Intermediary conclusion

In this part we have discussed the theories linking metacognition and inference, defined a metacognitive landscape where monitoring signals grow more complex together with cognitive and executive functions, discussed the social origins of self-regulation and implication towards legal responsibility and finally touched upon the ethical implication of this research beyond humans. From the mention of the role of the brain in learning to optimise fit between the agent and the environment, we have suggested and build upon the idea of procedural metacognition whereby metacognitive signals have a role, beyond the adjustment of learning, at correcting and adjusting behaviour to ensure this fit between the agent and both natural and social environments. Building on the metacognitive landscape (Fleming et al., 2012), we suggest that interdisciplinary research should aim at further defining the boundaries of and connctions between these dimentions in order to get a full picture of the computation and function of metacognitive signals. Overall, we summarise these various ares of research as defining metacogniton similarly to a thermostat for reliability. We illustrate this analogy in Fig. 3 where a simple binary thermostat tunes its control (Go- No-Go) depending on the monitored difference between the entropy of the observed system ( $\text{entropy}(\text{eau})=1$ ) and of its expected level of entropy ( $\text{entropy}(\text{ice})=0$ ). Similarly, looping back to our opening example of Truman as rational agent, we can illustrate his control over preserving his goal to escape the TV set instead of reverting back to the intuitive belief that this is the real word by comparing the reliability of the negative feedback he recieves to the expected level of reliability he expects from his beliefs ( $\text{reliability}(\text{movie set}=55\%$ ). In other word, this analogy suggests that metacognition adjusts the relibility of the deicison making itself by investing in proportional effective control such as by inhibiting impulse, switching strategy or coorrecting decisions. By simplifying the definition of rational agents to monitoring the gap between expected and observed (cognitivie and behavioural) reliability and in turn tuning effective control to ensure that behaviours follows through coherently (in spite of some accounted radomness

in the world), we turn to the research in cognitive neuroscience to review the ongoing research and discuss its contribution to the present definition of human metacognition as obtained from philosophy.



**Figure 3: Analogy of metacognition as a thermostat for reliability.** a. a simple binary thermostat of entropy: monitors the difference between its expected entropy ( $G=\text{entropy}(\text{ice})=0$ ) and its perceived entropy ( $P=\text{entropy}(\text{water})=1$ ). The monitoring signal causes the thermostat to invest energy into control to adjust the system below to match align with expected level of entropy. b. rational agent as a reliability thermostat: the agent perceive his world as a TV set and decides to act upon this belief but received negative feedback. While a simple binary model would update the agent's goal, the rational agent takes into account the expected reliability of his action's outcome to inhibit accordingly the intuitive reversal of intention to remain in the present environment

### 3. Cognitive neuroscience of metacognition

Operationalising the concepts of self-knowledge and reflection from their philosophical concepts into pragmatic methods and measures is an ongoing challenge. By testing these concepts on participants, empirical practices aim at shaping the multi-dimensional metacognitive landscape to refine it into a functional picture that can predict its successes and failures. The field of psychoanalysis developed by Sigmund Freud can be seen as the first set of methods for self-reflection. However the subjective narrative on which the method relies does not come with the required replicability and objectivity of modern science. The study of metamemory provided standard scales of evaluative judgments (feeling of knowing, tip of tongue...) to be tested on controllable set of material. However, the field of psychometrics providing frameworks to study the formation of subjective representations has shown powerful to identify the fundamental building blocks of confidence judgments. By providing cognitive representation tightly controlled by perceptual representation enabled to compare subjectively evaluated accuracy with its objective accuracy. However, defining a paradigm which will enable to study the formation of confidence levels in more realistic settings remains a central challenge in the field. In the remaining of this article, we refer to the study of metacognition as the study of confidence judgment in simple decision making tasks as it represents the majority of the ongoing research. Before doing so, we wish to stress the recent collaborative effort of researchers in the field of perceptual metacognition to agree on central questions to address in priority (Rahnev et al., 2021). Amongst these priorities are flagged the questions of: extending our metacognitive computation in simple perceptual tasks to more complex and ecologic decisions (c.f. part 3.1); the identification of the factors shaping metacognitive ability at the individual level (c.f. part 3.3); and the role of these metacognitive monitoring in conscious perception and behaviours (c.f. part 3.2).

### **3.1. Computation of confidence**

We discussed the theoretical work that paints metacognition as a thermostat monitoring the reliability of cognitive processes to subsequently adjust behaviour and effort. Focusing on this full loop from monitor to control, we here discuss this first side of the coin by highlighting how confidence measures behavioural reliability as objective accuracy and beyond. We present the empirical work that investigates the foundations of our levels of confidence: how it tracks – or not – the reliability of our decisions. We start by presenting measures of metacognitive ability that evaluate the link between subjective judgment and objective accuracy (part 3.1.1), we then present the cues on which confidence appear to rely to form these judgments (part 3.1.2). We then open on broader models than the normative one (part 3.1.3) and review the literature on a pressing question: domain generality (part 3.1.4).

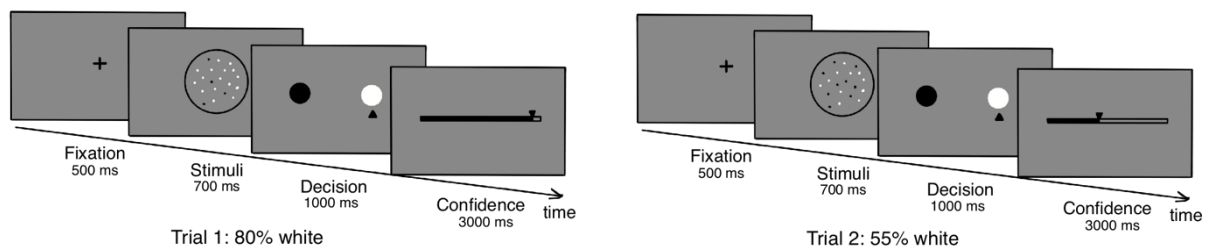
#### **3.1.1. Metacognitive accuracy**

##### *Normative measures and models of metacognitive accuracy*

If an agent's level of confidence in his choice appears to predict whether she will revise or invest in her choice, what shapes its truthfulness at capturing the decision's reliability? Do confidence levels make an accurate evaluation of a choice's reliability? How could our own brain capture subjectively such objective criteria? To test this question, defining a quantifiable measure of metacognitive accuracy as discriminating decision's accuracy with higher levels of accuracy than inaccurate ones is necessary.

To define such measure of the ability of metacognition to provide a subjective metric of objective decision reliability, psychometric has proven a fruitful framework by providing operationalizable tasks that relate objective signals with subjective precepts. In these tasks where participants are instructed to maximise accurate responses, an agent is said to be metacognitively accurate if she can distinguish her correct from incorrect decisions with respectively high and low levels of confidence. In other words, an agent has accurate metacognition if she can subjectively evaluate whether or not her own decision met the decision rule she intended to follow. To illustrate this relation between decision accuracy and confidence level, imagine you are foraging for mushrooms in the forest. An expert tells you that you will find two

types of mushrooms in the forest, both covered with black and white dots. He instructs you to pick the good ones which have a majority of white dots and to leave the bad ones which have a majority of black dots. Furthermore, having to collect a maximum of mushroom by night fall, you try to make these decisions within a couple of seconds for each mushroom. Here the decision rule you intend to follow is therefore to identify whether the mushroom presents a majority (*i.e.* criterion of >50%) of white dots or not. The first mushroom seems to present about 80% white dots, you take it, the second about 55% of white dots, you look once, then squint, and take it as well, and carry on at a steady pace. In such discrimination task, your confidence is expected to reflect your probability of being correct in your decision: and you should express higher confidence in having correctly taken the first mushroom than second, for which the evidence supporting its edibility was more conflicting.



**Figure 4:** Pair of trials for perceptual task with confidence report. Each trial presents successively on a computer screen a fixation cross, a stimulus with a mixture of white and black dots, a two-alternative forced choice (2AFC) and a confidence scale such as from “guessed” to “certain correct”. Reports are made under time pressure and confidence levels are generally observed to relate to the amount of evidence supporting the decision: higher in trial 1 than trial 2.

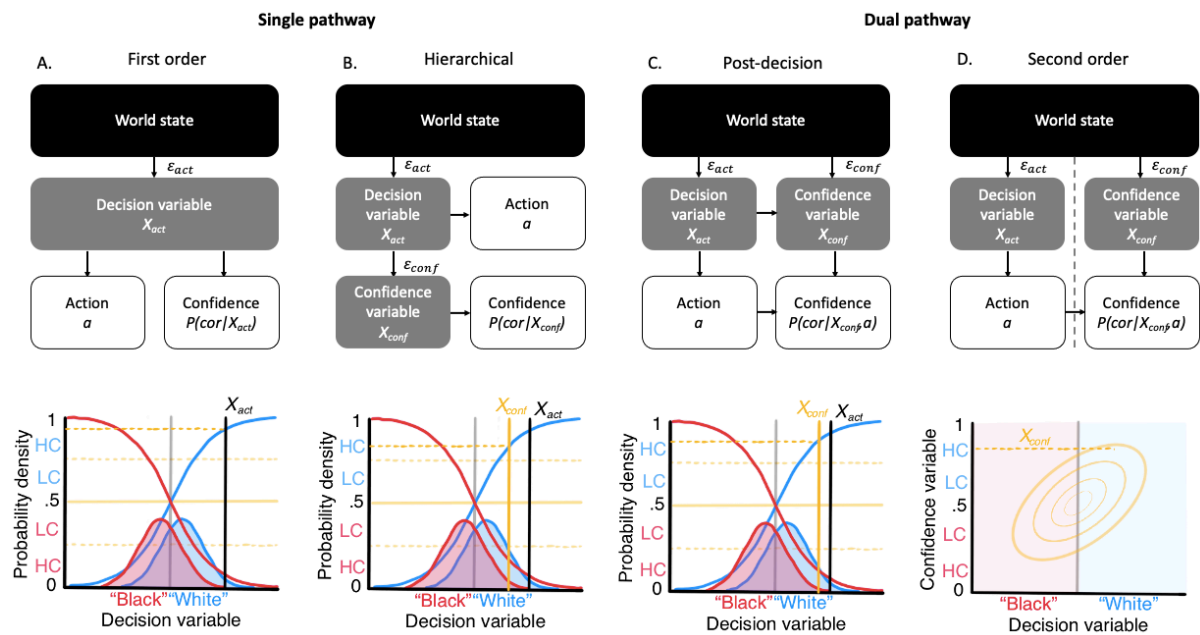
Relating to different assumptions about the workings of the metacognitive system (Fig. 4), different models were designed to capture and quantify metacognitive ability. We here present these models in two families as being or not strictly normative and discuss their limitations. First of all, parsimonious and strictly normative models capture metacognitive ability as the relation between confidence levels and decision accuracy. The recent collaborative effort in metacognitive

research compiled a rich data set (Rahnev et al., 2019) which confirmed this overall correlation between decision accuracy and confidence levels (Jin et al., 2021). These results therefore confirm that explicit levels of confidence provide participants with a subjective insight into the objective accuracy of their choices. This subject-to-subject correlation between average confidence and overall accuracy therefore captures the quality of a participant's insight, and thereby provides a metric to test the workings of confidence. However this most parsimonious model of confidence confounds key variables of the confidence computation that had to be captured and overcome (for review see Fleming & Lau, 2014). First of all, the development of second order signal detection theory (SDT2) enabled to distinguish two different metacognitive parameters: metacognitive sensitivity refers to the ability to distinguish with high and low confidence respectively accurate from inaccurate decisions; whereas metacognitive bias corresponds to the tendency to over or under estimate one's performance. Secondly, the effects of behavioural performance (level 1) on metacognitive sensitivity (level 2) were captured by new methods: adjusting participant performance technically (i.e. staircases) and normalising both measures together (i.e. M ratio). Nonetheless it is important to notice that the use of these artificial normalisations of performance were recently flagged to eventually distort the measurement of genuine metacognitive sensitivity (Guggenmos, 2021; Rahnev & Fleming, 2019).

While a finer Bayesian measure was since developed to best capture metacognitive sensitivity (Fleming, 2017), it is important to notice the limitation of these measures. First of all, these normative measures assume a stable difficulty all along the experiment on which the participant anchors criteria for confidence monitoring. Secondly and most importantly, these measures of metacognitive sensitivity aim at capturing the participant ability to track subjectively the objective accuracy of her decisions, as defined by laboratory instructions. This assumption of a normative monitoring of accuracy was recently flagged for the interpretations it draws. As for other normative models of decision making such as expected utility, Shekhar & Rahnev (2020) addressed that systematic dissociations from the norms were not to be taken as "metacognitive inefficiencies" but should be accounted as contributors to the model of confidence computation above and beyond this norm of accuracy.

These broader models of metacognition will be discussed in part 3.1.3 after defining the candidate cues on which confidence relies to seemingly track objective accuracy in perceptual task.

In the following part (3.1.2), we therefore focus on an alternative methods of investigation of metacognition not relating confidence to objectively measured accuracy but to candidate cues on which cognition could base this computation in order to infer its own accuracy. Building on signal detection theory, we start with perceptual predictors of the stimuli that are local (single pathway models Fig. 5a,b) and then extend on perceptual predictors of the more complex dual pathway models (Fig. 5c,d, for reviews see (Fleming & Daw, 2017; Maniscalco & Lau, 2016). We then promptly present the cognitive predictors on which confidence levels appear to rely to predict decision accuracy.



**Figure 5:** Normative models of perceptual confidence. These models (top part) illustrate candidate relation between decision accuracy and confidence levels (bottom part) and were adapted from Fleming & Daw, 2017 and Maniscalco & Lau, 2016. Single pathway models (A-B) assume that metacognitive information is limited to or derived from cognitive information. Further than the cognitive noise ( $\epsilon$ ), the hierarchical model (B) provides a metacognitive read-out noise ( $\epsilon$ ) that dissociates the quality of the decision (relating to variable  $X_{act}$ ) from confidence levels (resulting from  $X_{conf}$ ). Dual pathway models account for metacognitive evidence to be acquired

independently from the cognitive evidence (either combining both (C) or simply correlating (D)) and accounting for other sources of evidence such as post-decision motor action.

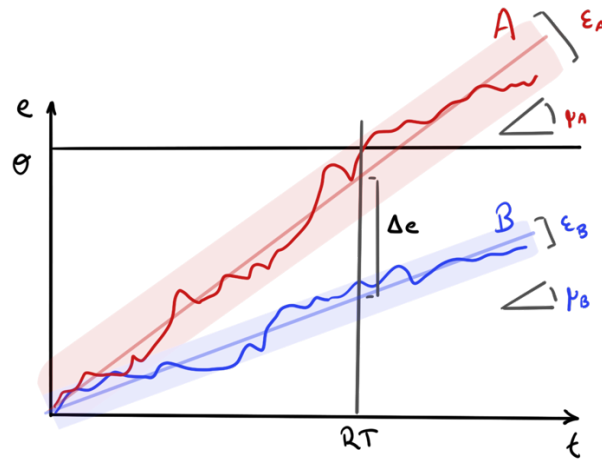
### **3.1.2. Predictors of confidence**

#### *Input reliability in single pathway*

The normative model of metacognition in perceptual tasks is supported by the consistent findings that confidence levels can be predicted by the reliability of perceptual evidence. Throughout various studies, two central parameters of evidence reliability seem to predict how our subjective levels of confidence might capture the objective accuracy of our decision: evidence strength and noise. In our example above with the mushrooms identification, signal strength equates to the distance between the criterion (50%) and evidence for the mushroom to be either mainly white or black (e.g. 80% white). Signal noise instead could be represented by the accessibility of this evidence: if the dots are small, far or partly covered. Overall, it was observed that while both parameters inform subjective confidence about the decision reliability, noise appears to be a better predictor than strength (Boldt et al., 2017; Spence et al., 2016). Furthermore, the relative contribution of both these cues was demonstrated to be subjective and maintained over time (De Gardelle & Mamassian, 2015; Navajas et al., 2017) but also to be encoded by distinct cerebral areas (Bang & Fleming, 2018). Therefore, as a subjective fingerprint, we observe that participants have their own sensitivity to different sources of stimulus reliability to monitor the accuracy of their choices.

The sequential sampling models elegantly account for these parameters of input reliability while also providing a proxy for confidence level as the difference between the evidence accumulators of each item.





**Figure 6: race-model of two competing items.** For both accumulators A and B, evidence  $e$  is accumulated with a drift rate  $\mu$  and a noise  $\varepsilon$  until the first one to reach the bound  $\theta$  is to be chosen. Confidence is then estimated as the difference of evidence  $\Delta e$  between both accumulators at the moment when bound is reached: response time  $RT$ . Evidence can continue to accumulate in dual pathway models (as in the post-decision model Fig 5c) and can provide additional evidence on which to build confidence level. This model also applying in value-based decision model (where evidence is retrieved for memory) can also account for a factor of attention  $\alpha$  that can modulate drift rate (equation 1). This attention boosting the evidence sampling is itself suggested to be driven by the value attributed to items. In turn, confidence level in a decision can influence the cognitive effort (e.g. attention or bound of required evidence) in order to adjust the performance to the context or goal.

Equation of the cognitive evidence accumulation for one alternative option as modulated by attention:

$$e_{t+1} = e_t + \alpha * \mu + \varepsilon \quad (1)$$

We can note that this equation also applies to value-based decisions where participants have a clear visual identification of the options but must sample from their memory the value they attribute to each option. For instance one might be presented with two ice creams (e.g. vanilla and chocolate) and gather how much one appreciates both flavors until one of both reach the sampling bound (Fig. 6)

### *Input reliability in dual pathway*

Different normative models have been proposed (Fig. 5) and while some seem more parsimonious than other, they predict different relation between confidence levels and evidence strength (e.g. Fig. 5: confidence variable dissociates from decision variable, Fleming & Daw, 2017). These predictors of confidence all together inform us about the computational nature of confidence levels. These predictors go beyond the previous “locus evidence”, and suggest a dual pathway models where further sources of evidence can inform confidence levels in order to estimate objective accuracy. By proposing a partly independent access to evidence than the decision-making system, these dual pathway models offer an explanation to a central question about metacognition: how can an agent both make a decision and also manage to know it is inaccurate? Besides the bridges between confidence levels and error detection, we present two candidate sources of evidence accumulation that could therefore explain how metacognition managed to evaluate the accuracy of a decision.

First of all, the sensitivity of confidence levels to unconscious cues suggest that metacognition has an independent access to perceptual evidence from the decision-making system (Charles et al., 2013; Kanai et al., 2010; Kunimoto et al., 2001; Meuwese et al., 2014; Rausch et al., 2018; Vlassova et al., 2014), a phenomenon that is also studied clinically as “blindsight” (Ko & Lau, 2012). It is interesting to notice that the early study of metacognition with Tip Of Tongue (ToT) phenomenon and Feeling of Knowing (FoK) was based on this concept that we could know whether some knowledge was present in our mind without having conscious access to it. (B. C. Smith, 2014). Functionally, this independent access of metacognition was demonstrated to enable agents to infer complex reward contingency from unconscious cues (Cortese et al., 2020).

Secondly, the privileged access of metacognition to perceptual evidence also enables it to access evidence after the decision is made, and so in two manners: by sampling further from the environment, or by recalling evidence from memory (Moran et al., 2015; Navajas et al., 2016a). These “post-decision models” (Fig. 5c) are

central to understand the possible overlap between error detection and the computation of confidence levels.

Whether this privilege of metacognition concerned perceptual evidence that is recalled, unconscious or even from other sources such as motor, these “dual pathway” models of metacognition suggest its reliance on working memory. We now discuss the cognitive predictors of confidence.

### *Cognitive reliability*

By accounting for the independent sampling of the metacognitive system, the second order model enables to better explain the cues on which the brain may rely to estimate the probability of our choices to be correct. Amongst these markers of cognitive reliability, error detection signals such as Error Related Negativity (ERN that accounts for conflicting perceptual and motor evidence) or error positivity (Pe) are strong predictor of respectively probability of an error to occur and its explicit detection (Boldt & Yeung, 2015). Other cognitive markers that predict both accuracy and its detection include eye tracking such as gaze shift frequency (GSF) and attention attribution (Sepulveda et al., 2020). Lastly, and aligning with evidence strength and attention attribution, the best known predictor of both confidence levels is response time that also tend to predict accuracy (Kiani et al., 2014; Patel et al., 2012).

While these markers of cognitive reliability both predict the probability to be correct and of high confidence, it is important to note that it does not necessarily suggest that metacognition tracks decision accuracy per se, but could instead appraise the reliability of cognitive processes themselves. While normative models provide useful assumptions to study the computation of confidence levels in psychometrics-like tasks, the systematic presence of factors deviating from this model are useful cues to reveal the genuine workings and function of metacognition.

### **3.1.3. Beyond normative models**

A recent review has highlighted the need to account for the systematic predictors of confidence which are often considered as “metacognitive inefficiencies” since they detach confidence levels from the probability to choose correctly. Here we review some metacognitive theories that attempt to provide a functional explanation for these consistently observed phenomenon.

#### **3.1.3.1. Local heuristics of cost benefit**

We define local predictors of confidence (as opposed to global) as the factors relating exclusively to the trial at hand (Rouault et al., 2019). As just reviewed, markers of cognitive reliability tend to cue confidence levels on the probability of the decision to be correct. However, a theory of “cognitive fluency” suggests that metacognition rather tracks the reliability of the decision-making process than the likelihood of success. Supporting this dissociation, the case of “metacognitive illusions” describes a phenomenon whereby the salience of evidence predicts response time and confidence levels while going against their probability of being correct (Koriat & Bjork, 2006; Rhodes & Castel, 2009).

As will be reviewed in part 3.2, this theory of cognitive fluency monitoring can suggest that confidence levels serve to adjust the reliability of the local decision-making process such as by correcting the made decision or supplying additional cognitive resources. However, further than simply adjusting the decision-making process in light of the single trial at hand, we now discuss possible theories for the tracking of decision reliability within a given context where learning can occur.

#### **3.1.3.2. Global monitoring**

Confidence levels reflect the reliability of the context in which decisions are made.

#### *Across trials consistency*

First, links between trials can be observed such as the called “confidence leak” where previous levels of confidence are observed to influence confidence in the present trial, independently from its accuracy (Rahnev et al., 2015; Rahnev & Denison, 2018). Another effect linking confidence to other trials is the observation that confidence

levels tend to be better predicted by the likelihood of choices to be repeated over time than their likelihood to be correct. This effect has inspired the Self-Consistency Theory of metacognition which suggests that this monitoring is mainly concerned with tracking the consistency of decisions rather than their accuracy (Koriat, 2012; Koriat et al., 2015; Koriat & Adiv, 2015). Initially, this theory attributes this effect to heuristic cues which are also shared with a group. However, confidence is also known to be more sensitive to evidence confirming one's choice than going against it (Peters et al., 2017; Samaha & Denison, 2022). Therefore confidence can be seen not only as tracking consistency but also to preferably sample from the cues that trigger consistency, thereby nudging behaviour into consistency through overconfidence. This theory can therefore be linked to the concept of self-determination which suggests a tendency of humans to be motivated by their own beliefs and behaviours in a positive feedback loop, independently from how they relate to the external world. The implication of building such confidence bias on behavioural flexibility and mental health will be discussed in part 2.3.3.

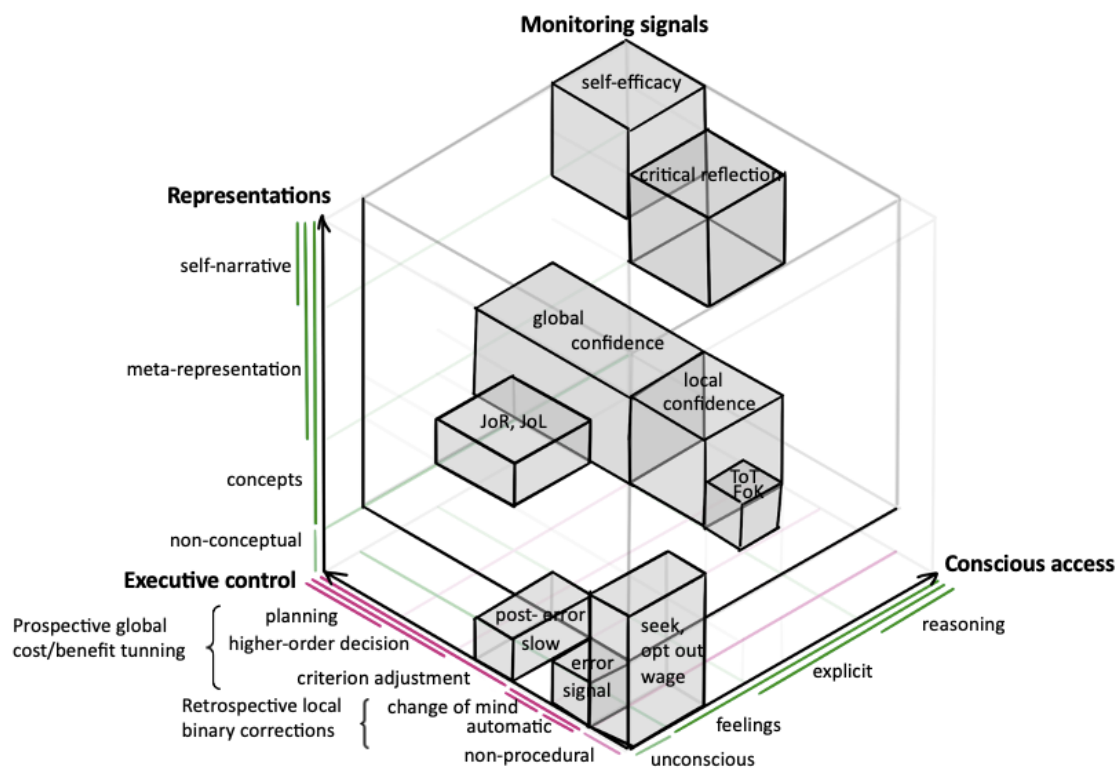
### *Contextual consistency*

Secondly, beyond reflecting tendencies across trials, confidence levels are known to reflect the regularities of the decision-making context: either as metacognitive learning from local confidence in absence of feedback (Guggenmos et al., 2016; Rouault et al., 2019), or as part of Bayesian learning (Bang & Fleming, 2018; Desender et al., 2019). In the latter, confidence levels in previous decision appeared to be better predictor of the global context's outcome than of the local trial difficulty. In perceptual tasks, this global representation of the task's reliability appeared as a useful signal to track the possibility of a change of context (Lee et al., 2021; Rouault et al., 2019). This evidence therefore suggests an ubiquitous role of monitoring signals in tracking the reliability of contexts further than local accuracy. Furthermore, these models of confidence linking learning and metacognition align with the opening part of the present review, shedding light on the workings of the brain's central function (*c.f.* part 2.1.1.).

We suggest that the study of value-based metacognition, from the economic side with learning about lotteries (rather than from preferential choices), provides a solid

framework to study the computation of confidence levels (Boldt et al., 2017; Hertz et al., 2018). As we are about to discuss (*c.f.* part 2.3), this tracking, not of the local probability of a decision to be correct, but of the reliability of a context or a lottery or item value is central to another type of behavioural flexibility: deciding how to decide, or meta-reasoning (Ackerman & Thompson, 2017; Boureau et al., 2015; Lieder et al., 2019).

From this contextual learning, metacognition was demonstrated to improve its calibration and become more sensitive (Chen et al., 2019; Rademaker et al., 2012; Sherman et al., 2015). Additionally, higher levels of declarative knowledge are formed such as with more general estimate of self-efficacy. This metacognitive tuning matured through experience and feedback is believed to take place throughout childhood to adulthood and to then continue to be adjusted throughout the agent's lifetime (Roebbers, 2017). Therefore, there literature suggests a strong bidirectional influence of learning on confidence and confidence on learning.



**Figure 7: The landscape of metacognitive signals between philosophy and cognitive neuroscience** (adapted from Fleming et al., 2012) Building on the previously defined landscape, cognitive neuroscience adds additional metacognitive signals to the picture: error signals such as ERN and Pe which lead to automatic adjustment of the decision process. Post-error slowing down enable to adjust the speed accuracy criterion to optimise future decisions after an error. Global confidence is learned contextual reliability and enable rational flexibility of decision rule between contexts. Self-efficacy also informs the agent about reliability of her previous as learned through experience. While the causality between the executive function and the monitoring signals varies, different tasks associate various monitoring signal(s) with executive controls.

### 3.1.4. Domain generality

We discussed the various cues on which confidence levels are based to infer the reliability of decisions and provide a potential common currency to eventually guide executive control. A popular question in the field of metacognition being debated is

the question of domain generality: does a same computation evaluate all types of decisions or does that depend on the domain in which they are performed?

While a large amount of research has aimed at investigating the possibility of a central monitoring system that could enable to use confidence as a common currency of reliability to trade between different types of evidence, decisions or beliefs, more recent literature points a finger at the common correlational approach which has limited explanatory power (Fleming, 2023). Indeed, as discussed in part 2.3.2, the tuning of metacognition to different tasks and contexts throughout development could be seen as a likely advantage to make rational decisions (as defined in Fig. 3). A meta-analysis of the recent work on that end tends to support this view: while some significant similarities are found in metacognitive ability throughout perceptual tasks, the link between confidence and accuracy does not seem to be the carried to memory tasks within individuals (Rouault, McWilliams, et al., 2018).

### **3.1.5. Intermediary conclusion.**

In light of our definition of metacognition as “a thermostat of reliability” as we suggested was given by philosophy (Fig. 3), the literature in cognitive neuroscience which focus on confidence levels informs us greatly about the computation of these monitoring signals. We note however that the present literature still presents an important gap between the tasks operationalised in the laboratory and presenting a functional role for behaviour (Fleming, 2023; Rahnev et al., 2021). We suggested that merging the research on confidence with the field of economics and value-based decisions could provide a useful bridge to resolve this gap in our metacognitive landscape (Fig. 6). Furthermore, this literature on the computation of monitoring signals tends to detach itself from the executive side of metacognition. To address this other side of the thermostat, we now review the empirical literature relating to the link between executive functions and metacognition.



### **3.2. Functions of confidence**

What could a robot or non-human animal gain with explicit procedural metacognition? While we discussed in length the models and contributors to the monitoring of such function, we now address the end of the tail, the other side of the coin: what executive controls can metacognition actually causally influence? Here we look at metacognition through the eye of executive control and discuss the various monitoring it relies on. Would computer that turns on its own fan to cool down its process and ease its processing be metacognitive? Does Truman (in *The Truman Show*, 1998) who makes up some plans to escape a world he believes to be staged present some metacognitive traits? Building on our metacognitive landscape as previously defined by philosophy and the science of metacognitive monitoring, we now aim at shredding the picture further by reviewing the empirical literature that actually closes the loop on procedural metacognition. We divide this part into four based on executive control. First we go over non-procedural report of uncertainty that have no effect beyond the decision. Secondly we present retrospective executive control that aim at adjusting at correcting a past decision. The last part concerns prospective executive functions that use contextual learning to either adjust a decision criterion, switch strategy even engage in sophisticated behaviour such as planning and using extended cognitive resources to meet one's goal.

#### **3.2.1. Non-procedural monitoring**

Uncertainty monitoring can appear to be ubiquitous to any decisions regardless of its effect on behaviour or the agent who chooses. This monitoring can be communicated in various ways without appearing to affect behaviour, past or future decisions. In non-verbal animal literature, the ability to opt out of difficult decisions can be argued to be a sophisticated behaviour. However, as for conscious access itself, these decisions were suggested not to rely on explicit and conscious access but instead to rely on associative learning (Owen et al., 2006; J. D. Smith et al., 2012). By wagering on the decision such as by investing time by waiting for a reward or instead spending time looking for more evidence before making the decision, simple behaviours can reflect the reliability of the decision. These behaviours where

demonstrated to rely also on evidence ambiguity further than on simple reward association, thereby appearing as a possible root of the subjective monitoring of one's decision reliability (Kepecs et al., 2008). Other non-procedural monitoring signals can be found in explicit reports such as Tip-of-Tongue or Feeling-of-knowing which can serve to communicate likelihood of knowledge retrieval but do not guide learning like "Judgments of learning" would. Overall these non-procedural monitoring signals suggest that the metacognitive monitoring of decision reliability is ubiquitous and does not depend on the opportunity to correct or adjust one's behaviour.

### **3.2.2. Retrospective local corrections**

Retrospective monitoring signals (i.e. evaluating an already made decision) were suggested to predict change of mind. This corrective behaviour is largely observed in studies done under time pressure to ensure the presence of errors and provide ground for their detection. In such paradigms, participants are asked to report how confident they are that their decision maximises the decision rule: whether it be about the side towards which a Gabor patch seems to be rotated towards, or the snack for which they would be the most willing to pay. In such studies, confidence levels were found to predict change of mind (De Martino et al., 2013; Folke et al., 2017). However, in a simple laboratory task where some trials' confidence levels were boosted by evidence strength (while keeping performance equal), results failed to conclude on such a causal role of confidence on behavioural correction (Koizumi et al., 2015). Furthermore, as will be discussed in part 3.3.3, the degree of control one possess upon her behaviour is not necessarily predicted by the ability to detect incorrect decisions.

These results could suggest that metacognition is not cut out to track the optimality of a decision in the maximising sense intended by researchers in the laboratory. For instance, referring back to the Self-Consistency Theory which we introduced above (*c.f.* part 3.1.3.2.), confidence levels were suggested instead to rely on heuristic cues to track choices that are the most likely to be repeated without necessarily being the cause for this behavioural correction. In other words, certain cues could predict both choice consistency and confidence without them being causally linked. According

to this theory metacognition would thereby monitor (and inform the agent about) decision reliability in the sense that it would be most likely to be stable across time and agents, but being detached of the maximising behaviour assumed by our descriptive models.

Together, these results on confidence could suggest that independently from the ability of metacognition to track the optimality of our behaviour, our executive system might not obey such normative models. Two conclusion can be drawn: First, there is a functional gap between metacognitive monitoring and control: while being possible catalysts of behavioural correction, monitoring signals might not be sufficient to initiate it. Secondly, metacognitive monitoring signals might not track rational norms of behaviour as maximising (or not) the decision rule. Instead monitoring signals could provide the agent with a more “bounded rational” evaluation on a rather continuous scale, differentiating various levels of sub-optimality as requiring correction or not. In other words, metacognitive monitoring could help tune the amount of executive effort to be invested in order to keep behaviour at a desired degree of success: a kind of reliability thermostat (Fig. 3). If metacognition would therefore aim at ensuring a subjective notion of success rather than a normative one, one could expect a variation amongst contexts (e.g. value domain or task) and agents. In other words, metacognition could be studied not for its ability to accurately detect binary errors, but for the degree of success to which it is sensitive (i.e. the thermostat criterion). We suggest that, beyond psychometric tasks which are modelled on the binary notion of accuracy, developing economic tasks where items are encoded on continuous scales of subjective value could be a strong foundation to investigate such pragmatic function of procedural metacognition. In a nutshell, if monitoring signals serve for executive control, then rather than assuming a normative notion of accuracy, a more ecological notion that accounts for the cost of this output control should be established, therefore framing metacognitive monitoring at the centre of a bounded rational framework.

### **3.2.3. Prospective global regulations**

In this part, we present the multiple facets of metacognition as a thermostat of cognitive and behavioural reliability. Stepping beyond the normative models of metacognition as tracking the correctness of the past decisions, we review here the research investigating how monitoring signals tune the decision process in order to adjust the level of effort required to make reliable decisions.

#### *Prospective local*

At the local level, confidence levels in past decisions appear to regulate the decision-making process of future decisions by optimizing the fit of the speed-accuracy criterion in accordance to the context difficulty. With dynamic decision models as presented above (e.g. race model, Fig. 6) this adjustment can be modelled by increasing the decision bound (i.e. increase response time) in order to keep the desired level of performance. This criterion adjustment can also be seen to push participants to seek additional information or even to invest and pay for more information in order to reach their desired level of performance (Baldson et al., 2020; Desender et al., 2018, 2019; Schulz et al., 2021). Even in isolated local decisions, empirical results therefore suggest that the decision-making process is adjusted to maintain a desired performance.

The field of reasoning study the regulation of the decision making process within the decision process, before the decision is made. Similarly, feelings of rightness and intermediate levels of confidence were suggested to redirect the decision towards the alternative option before the decision was made, often enabling the selection of the optimal option against the initial intuitive option with lower value (Ackerman & Thompson, 2017).

Therefore, either within a decision making process or between two successive local decisions, monitoring signals appear to adjust the cost of the decision (e.g. increase attention, seek more evidence...) in order to maintain performance. At the local level, metacognition can therefore be seen as a reliability thermostat by adjusting level of effort to the difference between desired performance and predicted performance.

### *Prospective Global*

We discussed before (c.f. Part 3.1.3.3) that metacognition seems to ubiquitously upcoming decisions. We then just reviewed how confidence appears to adjust the dynamic decision-making process to ensure a certain degree of reliability (probability correct) relative to effort attributed in a bounded rational framework.

At the global level,

We previously discussed (c.f. Part 3.1.3.3) how confidence levels are linked to learning by predicting the reliability of a decision in a given context. Reciprocally, confidence levels have also been demonstrated to tune the learning process by weighting the update from local feedback or confidence with their expected probability to occur (Guggenmos et al., 2016; Salem-garcia et al., 2021). A neurofeedback study also demonstrated participants's ability to accurately predict their accuracy with their confidence levels and learn to improve their performance in a 2AFC task where the stimuli actually were subliminal (Cortese et al., 2020). In other words, the learning is adjusted by accounting for expected contextual reliability.

When contexts are volatiles, this tracking of global reliability was suggested to serve the pragmatic function of fitting the switch from one strategy to another in order to ensure that decisions meet the agent's goal (Boureau et al., 2015). Indeed, in such new contexts come the dilemma of efficiently switching from explore to exploit: when the agent notices that the actual output from local decisions does not match expected output from the strategy in the context, while accounting for the reliability of the strategy, then the context must have changed. In Truman's example, neighbors are expected to go about their days freely and stochastically (expected low reliability) but when he notices that they are going around the block on a loop (observed high reliability) then he concludes not to be in the real world but on a TV set, and switches from the strategy of exploiting the environment to exploring beyond known territories. However the literature is split whether this task of "deciding how to decide" (F. Becker & Lieder, 2021; Boureau et al., 2015) since some researchers conclude that this higher-order decision does not require metacognitive monitoring to be explained (Erev & Barron, 2005; Lieder & Griffiths, 2017).

When facing new or counter-intuitive contexts, the habitual strategy can dissociate from the goal-directed one. In that case, the agent ought to fit the flexibility criterion between exploiting and exploring both strategies. When this adjustment is done within the decision process (before the decision is made), continuous monitoring of the prospective decision can cue the agent about the sub-optimality of the fast and automatic response and provide the opportunity to overcome this heuristic by switching to the goal-directed strategy instead. While reasoning requires an analytic process to select the correct decision, meta-reasoning relies upon acquired knowledge (e.g. about the reliability of strategies, self efficacy) and relies heavily on working memory and attentive awareness (Ackerman & Thompson, 2017). In other words, as suggested by theories of metacognitive development, meta-reasoning supports rational behaviour by overwriting heuristics to follow one's goal. This process relies on the metacognitive ability to monitor sub-optimal decision-making and use metacognitive knowledge to catalyse this goal-directed behaviour. Since metacognition can be trained (c.f. part 3.3.3.), goal directed behaviour was suggested to be trainable together with metacognition (Lieder et al., 2018; Lieder & Griffiths, 2017). Thereby, in education and in the professional environment, metacognition was suggested to predict the ability to adapt efficiently to new environment by learning and adjusting behaviour faster (Fleming, 2021).

Beyond the ability to overwrite heuristic decisions can rely on perspective taking and overcoming temporal or social discounting. In these difficult decisions, metacognition was also demonstrated to predict the ability to predict sophisticated behaviour: the present shredding of future option to nudge goal directed behaviour and prevent heuristic decisions (Soutschek et al., 2021).

All together the literature linking metacognitive monitoring and control suggests a tight overlap with the learning mechanism which together tune (as a thermostat) the cost one ought to invest in the decision process in order to thrive towards one's goal. In a nutshell, these results suggest that some of the limitations encountered in the research on confidence levels in psychometric tasks (c.f. part 3.1, 3.2) could be best answered by reframing this research in value-based (economics and learning) and

reasoning tasks. In the following part, we discuss in more details how the evolution of metacognition appear to go hand in hand with higher levels of agency (c.f. part 3.3.1.), and how learning and education are critical for metacognitive training in human development (c.f. part 3.3.2.). Lastly, we discuss the link between metacognitive deficiencies and psychopathologies (c.f. part 3.3.2.).

### **3.3. Subjective Factors of metacognitive ability**

#### **3.3.1. Evolution of metacognition**

Rather than an open review of the literature on non-human animal metacognition (c.f. part 2.4.) that often aims at testing the neurobiology of implicit monitoring signals, we suggest reframing the present question in a functional theory. Aligning with the view of the brain's role in learning to adjusting the fit of the agent with the environment (c.f. part 2.1.), a recent book by Tomasello (2022) develops a theory of evolution as providing increasing levels of agency. The author suggests that by developing increasingly advanced models of the natural environment, animal species have learned to optimise their behaviour to navigate efficiently and even rationally while following their own goals. By developing social skills and conceptual language humans have then developed advanced models of a social environment comporting social norms detached from the natural environment. These increasing levels of agency are seen as degrees of freedom for the species to flexibly adapt to new challenging environments. More specifically, the author categorises 4 levels of agency, the latter being unique to humans with conceptual language defining an intricate collective environment relying on social norms. Arguably, the social role of explicit levels of confidence evolved as supporting this fourth level of agency (c.f. part 2.2.). is central to the notion of legal responsibility (c.f. part 2.3.).

We suggest that this functional framework could be most appropriate to develop an evolutionary model of procedural metacognition. Indeed, bridging theories of metacognition relating to both learning (c.f. part 2.1.) and goal-directed behaviour (c.f. part 3.2.3), the metacognitive landscape (Fig. 7) could best be structured by linking different cognitive processes to monitoring signals and executive function, pruning our model to provide new empirical hypothesis to be tested together with a

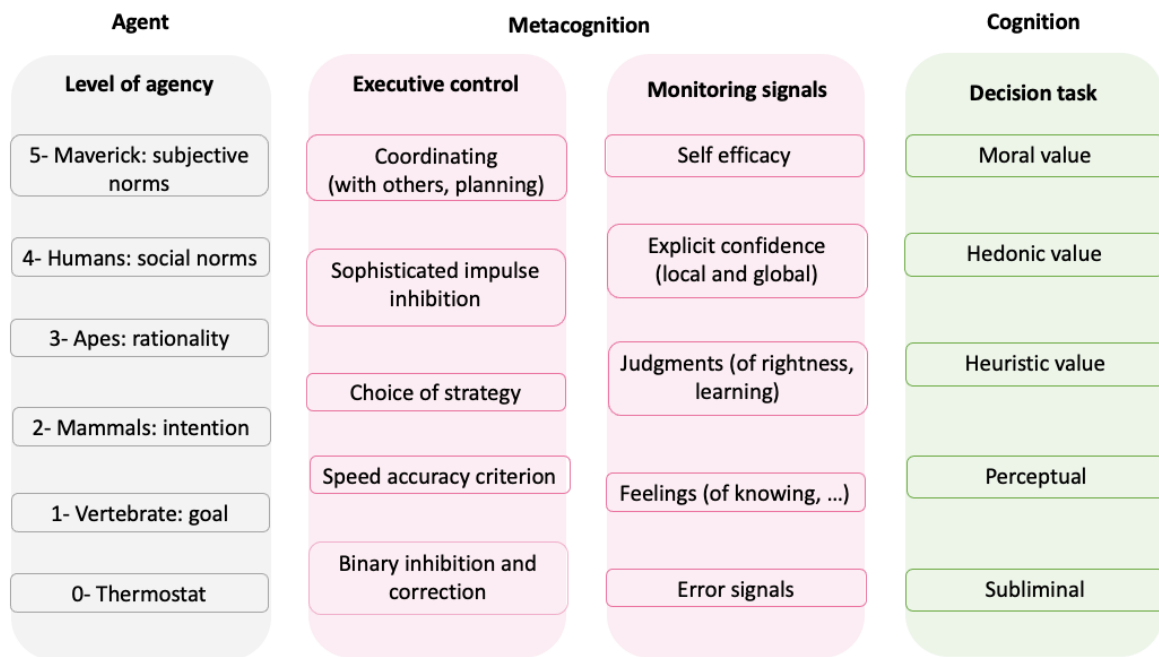
pragmatic and unified theory of the brain and metacognition as a reliability thermostat for agency.

### **3.3.2. Development of metacognition and metacognitive enhancement**

The development of metacognition, with the tuning of monitoring signals to provide responsible autonomy, takes place from childhood to adulthood and continues to evolve all along adult life. It is suggested that this ability to self-evaluate comes from education and the providing of feedback from parents revealing cultural differences (Roebbers, 2017; Weil et al., 2013). Since metacognition is described as a predictor of ability to learn and be successful professionally, it was suggested the school curriculum should emphasis the development of auto evaluation and its tuning (Fleming, 2021; Gilbert et al., 2019).

We suggest that the development of advanced abilities at critical analysis and self-evaluation could be key to developing subjective value and beliefs partly independently from social trends and norms. We suggest that such heightened levels of self-regulation could catalyse the emergence of a sense of personal identity in young adults. Supporting a high level of autonomy, as an additional level of agency, we suggest a possible additional level of agency as relying on this advanced metacognitive skill which be entitled maverick as reference to our opening example of Truman (Fig 8). While a study demonstrated that extreme political views are often accompanied by low metacognitive ability (Rollwage et al., 2018), research investigating the training of metacognitive ability could certainly support the development of well-rounded and autonomous agents (F. Becker & Lieder, 2021; Carpenter et al., 2019).





**Figure 8: Thermostats for agency: the computation and function of metacognitive monitoring signals.** Adaptation of the metacognitive landscape (Fig 7) and metacognitive thermostat (Fig 3) into a linear model whereby levels of agency of an agent (grey) rely on executive functions which rely on monitoring signals, enabling the agent to perform in various decision-making tasks. By refining the definition of, and connection between the above elements, research could provide an integrated theory of procedural metacognition and propose new hypotheses for its computation and function.

### 3.3.3. Psychopathologies and metacognitive therapies

Recent research has suggested a link between metacognitive bias and psychological disorders such as anxiety, social withdrawal and obsessive compulsive disorders (OCD) (Rouault, Seow, et al., 2018; Seow et al., 2021). While metacognitive ability to distinguish correct from incorrect decisions does not seem significantly involved, this global trait could be linked to a deficiency in metacognitive development and calibration. Furthermore, since self-efficacy is seen as an important factor for monitoring signals to recruit metacognitive control (Ackerman & Thompson, 2017), and that under-confidence is linked to self-efficacy, these results suggest that the recalibration of monitoring signals could contribute to rehabilitating these populations. Clinicians have also attempted to incorporate metacognition into a

therapeutical method by harnessing its potential for self awareness (Lysaker et al., 2018).

#### **4- General Conclusion**

Throughout philosophy and cognitive neuroscience, the term of “metacognition” is used as an umbrella term to investigate various mental processes and behaviours. Here we presented a landscape for procedural metacognition as defined by both fields. Mainly, we aim at positioning different types of metacognitive monitoring as serving different levels of executive control, thereby painting metacognition as a thermostat for bounded rationality. We suggest that this monitoring of reliability relies on an integration of various elements such as globally the reliability of the decision rule in fulfilling the goal; and locally of the decision to fulfil the decision rule (i.e. coherence in normative terms). We present empirical evidence supporting the computation of confidence levels and their effect on behaviour and discuss how reframing the research on metacognition with value-based tasks could help establish an integrated model for human agency.

## References

- Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, 21(8), 607–617. <https://doi.org/10.1016/j.tics.2017.05.004>
- Allen, M., Frank, D., Samuel Schwarzkopf, D., Fardo, F., Winston, J. S., Hauser, T. U., & Rees, G. (2016). Unexpected arousal modulates the influence of sensory noise on confidence. *ELife*, 5(OCTOBER2016), 1–17. <https://doi.org/10.7554/eLife.18103>
- Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., & Frith, C. (2012). What failure in collective decision-making tells us about metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1350–1365. <https://doi.org/10.1098/rstb.2011.0420>
- Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, 329(5995), 1081–1085. <https://doi.org/10.1126/science.1185718>
- Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, 11, 1–11.
- Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 115(23), 6082–6087. <https://doi.org/10.1073/pnas.1800795115>
- Becker, F., & Lieder, F. (2021). Promoting metacognitive learning through systematic reflection. *Workshop on Metacognition in the Age of AI. Thirty-Fifth Conference on Neural Information Processing Systems, NeurIPS*. <https://doi.org/10.13140/RG.2.2.24598.88642>
- Bennett, C. (2016). The Role of Remorse in Criminal Justice. *Oxford Handbooks Online*. <https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935383.001.0001/oxfordhb-9780199935383-e-37>
- Beyer, F., Sidarus, N., Fleming, S., & Haggard, P. (2018). Losing control in social situations: How the presence of others affects neural processes related to sense of agency. *ENeuro*, 5(1), 1–13. <https://doi.org/10.1523/ENEURO.0336-17.2018>
- Bigenwald, A., & Chambon, V. (2019). Criminal responsibility and neuroscience: No revolution yet. *Frontiers in Psychology*, 10(JUN), 1–19. <https://doi.org/10.3389/fpsyg.2019.01406>
- Bo, S., Kongerslev, M., Dimaggio, G., Lysaker, P. H., & Abu-Akel, A. (2015). Metacognition and general functioning in patients with schizophrenia and a history of criminal behavior. *Psychiatry Research*, 225(3), 247–253.

<https://doi.org/10.1016/j.psychres.2014.12.034>

- Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 43(8), 1520–1531. <https://doi.org/10.1037/xhp0000404>
- Boldt, A., & Yeung, N. (2015). Shared neural markers of decision confidence and error detection. *Journal of Neuroscience*, 35(8), 3478–3484. <https://doi.org/10.1523/JNEUROSCI.0797-14.2015>
- Boureau, Y. L., Sokol-Hessner, P., & Daw, N. D. (2015). Deciding How To Decide: Self-Control and Meta-Decision Making. *Trends in Cognitive Sciences*, 19(11), 700–710. <https://doi.org/10.1016/j.tics.2015.08.013>
- Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, 148(1), 51–64. <https://doi.org/10.1037/xge0000505>
- Carroll, J. E. (2016). Brain Science and the Theory of Juvenile Mens Rea. *North Carolina Law Review*, 94(2), 539–600.
- Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, 73, 80–94. <https://doi.org/10.1016/j.neuroimage.2013.01.054>
- Chen, B., Mundy, M., & Tsuchiya, N. (2019). Metacognitive Accuracy Improves With the Perceptual Learning of a Low- but Not High-Level Face Property. *Frontiers in Psychology*, 10(July), 1–15. <https://doi.org/10.3389/fpsyg.2019.01712>
- Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press USA.
- Cortese, A., Lau, H., & Kawato, M. (2020). Unconscious reinforcement learning of hidden brain states supported by confidence. *Nature Communications*, 11, 1–14. <https://doi.org/10.1038/s41467-020-17828-8>
- De Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLoS ONE*, 11(1). <https://doi.org/10.1371/journal.pone.0147901>
- De Gardelle, V., & Mamassian, P. (2015). Weighting mean and variability during confidence judgments. *PLoS ONE*, 10(3), 1–11. <https://doi.org/10.1371/journal.pone.0120870>
- De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, 16(1), 105–110. <https://doi.org/10.1038/nn.3279>
- Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc. Natl. Acad. Sci. USA*, 95(November),

- 14529–14534. <https://doi.org/10.1111/j.1749-6632.2001.tb05714.x>
- Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed-accuracy tradeoff for subsequent decisions. *ELife*, 8. <https://doi.org/10.7554/eLife.43499>
- Desender, K., Boldt, A., & Yeung, N. (2018). Subjective Confidence Predicts Information Seeking in Decision Making. *Association for Psychological Science*. <https://doi.org/10.1177/0956797617744771>
- Erev, I., & Barron, G. (2005). On Adaptation , Maximization , and Reinforcement Learning Among Cognitive Strategies. *Psychological Review*, November. <https://doi.org/10.1037/0033-295X.112.4.912>
- Falk, A., & Szech, N. (2013). Morals and markets. *Science*, 340(6133), 707–711. <https://doi.org/10.1126/science.1231566>
- Fleming, S. M. (2017). HMeta-d: hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, 2017(1), 1–14. <https://doi.org/10.1093/nc/nix007>
- Fleming, S. M. (2021). *Know Thyself: The Science of Self-Awareness*. Basic Books.
- Fleming, S. M. (2023). Metacognition and confidence: a review and synthesis. *Annual Review of Psychology*.
- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, 124(1), 91–114. <https://doi.org/10.1037/rev0000045>
- Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: Computation, biology and function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1280–1286. <https://doi.org/10.1098/rstb.2012.0021>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8(JULY). <https://doi.org/10.3389/fnhum.2014.00443>
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1), 17–19. <https://doi.org/10.1038/s41562-016-0002>
- Fotopoulou, A., Rudd, A., Holmes, P., & Kopelman, M. (2009). Self-observation reinstates motor awareness in anosognosia for hemiplegia. *Neuropsychologia*, 47(5), 1256–1260. <https://doi.org/10.1016/j.neuropsychologia.2009.01.018>
- Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *The Journal of Philosophy*, 68(1), 5–20. <https://doi.org/10.2307/j.ctvvh84xp.15>
- Friston, K. (2011). Embodied Inference : or “ I think therefore I am , if I am what I think .” *The Implications of Embodiment (Cognition and Communication)*, January 2011, 89–125.

- Friston, K. J. (2018). Active Inference and Cognitive Consistency. *Psychological Inquiry*, 29(2), 67–73. <https://doi.org/10.1080/1047840X.2018.1480693>
- Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, 6(4), 187–214. <https://doi.org/10.1080/17588928.2015.1020053>
- Frith, C. D. (2012). The role of metacognition in human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1599), 2213–2223. <https://doi.org/10.1098/rstb.2012.0123>
- Gilbert, S. J., Bird, A., Carpenter, J. M., Fleming, S. M., Sachdeva, C., & Tsai, P. C. (2019). Optimal Use of Reminders: Metacognition, Effort, and Cognitive Offloading. *Journal of Experimental Psychology: General*, 1–56. <https://doi.org/10.1037/xge0000652>
- Greco, J. (2019). The Social Value of Reflection. *Philosophical Studies Series*, 141, 45–57. [https://doi.org/10.1007/978-3-030-18266-3\\_4](https://doi.org/10.1007/978-3-030-18266-3_4)
- Guggenmos, M. (2021). Validity and reliability of metacognitive performance measures. *PsyArXiv*. <https://psyarxiv.com/jrkzm/>
- Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *ELife*, 5(MARCH2016), 1–19. <https://doi.org/10.7554/eLife.13388>
- 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), 54–61. <https://doi.org/10.1016/j.cognition.2010.06.010>
- Henriksen, M. (2020). Variational Free Energy and Economics Optimizing With Biases and Bounded Rationality. *Frontiers in Psychology*, 11(November), 1–10. <https://doi.org/10.3389/fpsyg.2020.549187>
- Hertz, U., Bahrami, B., & Keramati, M. (2018). Stochastic satisficing account of confidence in uncertain value-based decisions. *PLoS ONE*, 13(4), 1–23. <https://doi.org/10.1371/journal.pone.0195399>
- Hertz, U., Palminteri, S., Brunetti, S., Olesen, C., Frith, C. D., & Bahrami, B. (2017). Neural computations underpinning the strategic management of influence in advice giving. *Nature Communications*, 8(1), 1–12. <https://doi.org/10.1038/s41467-017-02314-5>
- Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. *Neural Computation*, 33(2), 398–446. [https://doi.org/10.1162/neco\\_a\\_01341](https://doi.org/10.1162/neco_a_01341)
- Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences*, 24(5), 349–362. <https://doi.org/10.1016/j.tics.2020.02.007>

- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*(APR), 1–14. <https://doi.org/10.3389/fpsyg.2012.00096>
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Jin, S., Verhaeghen, P., & Rahnev, D. (2021). Moderators of relative confidence calibration: A meta-analysis of the across-subject relationship between confidence and accuracy. *PsyArXiv Preprints*. <https://doi.org/10.31234/osf.io/mhbj9>
- Kahneman, D. (2003). A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist, 58*(9), 697–720. <https://doi.org/10.1037/0003-066X.58.9.697>
- Kanai, R., Walsh, V., & Tseng, C. H. (2010). Subjective discriminability of invisibility: A framework for distinguishing perceptual and attentional failures of awareness. *Consciousness and Cognition, 19*(4), 1045–1057. <https://doi.org/10.1016/j.concog.2010.06.003>
- Kepecs, A., Uchida, N., Zariwala, H. A., & Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature, 455*(7210), 227–231. <https://doi.org/10.1038/nature07200>
- Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron, 84*(6), 1329–1342. <https://doi.org/10.1016/j.neuron.2014.12.015>
- Kluwe, R. H. (1982). Cognitive Knowledge and Executive Control: Metacognition. *Animal Mind — Human Mind*, 201–224. [https://doi.org/10.1007/978-3-642-68469-2\\_12](https://doi.org/10.1007/978-3-642-68469-2_12)
- Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1401–1411. <https://doi.org/10.1098/rstb.2011.0380>
- Koizumi, A., Maniscalco, B., & Lau, H. (2015). Does perceptual confidence facilitate cognitive control? *Atten Percept Psychophys, March*, 1295–1306. <https://doi.org/10.3758/s13414-015-0843-3>
- Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review, 119*(1), 80–113. <https://doi.org/10.1037/a0025648>
- Koriat, A., & Adiv, S. (2015). The self-consistency theory of subjective confidence. *Oxford Handbook of Metamemory, 1*(June), 1–25. <https://doi.org/10.1093/oxfordhb/9780199336746.013.18>
- Koriat, A., Adiv, S., & Schwarz, N. (2015). Views That Are Shared With Others Are Expressed With Greater Confidence and Greater Fluency Independent of Any Social Influence. *Personality and Social Psychology Review, 20*(2), 176–193. <https://doi.org/10.1177/1088868315585269>

- Koriat, A., & Bjork, R. A. (2006). *Mending Metacognitive Illusions : A Comparison of Mnemonic-Based and Theory-Based Procedures*. *32*(5), 1133–1145.  
<https://doi.org/10.1037/0278-7393.32.5.1133>
- Kunimoto, C., Miller, J., & Pashler, H. (2001). Confidence and Accuracy of Near-Threshold Discrimination Responses. *Consciousness and Cognition*, *10*, 294–340.  
<https://doi.org/10.1006/ccog.2000.0494>
- Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Current Biology*, *27*(6), 821–832. <https://doi.org/10.1016/j.cub.2017.02.026>
- Lee, A. L. F., de Gardelle, V., & Mamassian, P. (2021). Global visual confidence. *Psychonomic Bulletin and Review*, *28*(4), 1233–1242. <https://doi.org/10.3758/s13423-020-01869-7>
- Legrand, N., Engen, S. S., Correa, C. M. C., Mathiasen, N. K., Nikolova, N., Fardo, F., & Allen, M. (2021). Emotional metacognition: stimulus valence modulates cardiac arousal and metamemory. *Cognition and Emotion*, *35*(4), 705–721.  
<https://doi.org/10.1080/02699931.2020.1859993>
- Lieder, F., Chen, O. X., Krueger, P. M., & Griffiths, T. L. (2019). Cognitive prostheses for goal achievement. *Nature Human Behaviour*, *3*(10), 1096–1106.  
<https://doi.org/10.1038/s41562-019-0672-9>
- Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review*, *June*. <https://doi.org/10.1037/rev0000075>
- Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*(4), 1–27.  
<https://doi.org/10.1371/journal.pcbi.1006043>
- Logan, G. D., & Crump, M. J. C. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, *330*(6004), 683–686.  
<https://doi.org/10.1126/science.1190483>
- Lysaker, P. H., Gagen, E., Moritz, S., & Schweitzer, R. D. (2018). Metacognitive approaches to the treatment of psychosis: A comparison of four approaches. *Psychology Research and Behavior Management*, *11*, 341–351. <https://doi.org/10.2147/PRBM.S146446>
- Maniscalco, B., & Lau, H. (2016). The signal processing architecture underlying subjective reports of sensory awareness. *Neuroscience of Consciousness*, *2016*(1), 1–17.  
<https://doi.org/10.1093/nc/niw002>
- Metcalf, J., & Son, L. K. (2012). Anoetic, noetic and auto-noetic metacognition. *In The Foundations of Metacognition*, 289–301.
- Meuwese, J. D. I., van Loon, A. M., Lamme, V. A. F., & Fahrenfort, J. J. (2014). The subjective experience of object recognition: Comparing metacognition for object



- detection and object categorization. *Attention, Perception, and Psychophysics*, 76(4), 1057–1068. <https://doi.org/10.3758/s13414-014-0643-1>
- Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, 78, 99–147. <https://doi.org/10.1016/j.cogpsych.2015.01.002>
- Moulin, C., & Souchay, C. (2015). An active inference and epistemic value view of metacognition. *Cognitive Neuroscience*, 6(4), 221–222. <https://doi.org/10.1080/17588928.2015.1051015>
- Navajas, J., Bahrami, B., & Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences*, 11, 55–60. <https://doi.org/10.1016/j.cobeha.2016.05.005>
- Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P. E., & Bahrami, B. (2017). The idiosyncratic nature of confidence. *Nature Human Behaviour*, 1(11), 810–818. <https://doi.org/10.1038/s41562-017-0215-1>
- Nelson, T. O. (2000). Consciousness, self-consciousness, and metacognition. *Consciousness and Cognition*, 9(2 Pt 1), 220–223. <https://doi.org/10.1006/ccog.2000.0439>
- Nelson, Thomas O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, 26(C), 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Nisbett, R. E., Wilson, T. D., Kruger, M., Ross, L., Indeed, A., Bellows, N., Cartwright, D., Goldman, A., Gurwitz, S., Lemley, R., London, H., & Markus, H. (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review*, 3.
- Owen, A. M., Coleman, M. R., Boly, M., Davis, M. H., Laureys, S., & Pickard, J. D. (2006). Detecting Awareness in the Vegetative State. *Sciencemag*, 313(September), 2006.
- Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748), 4853–4860. <https://doi.org/10.1098/rspb.2012.1847>
- Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, 1(7). <https://doi.org/10.1038/s41562-017-0139>
- Pezzulo, G., Rigoli, F., & Friston, K. J. (2018). Hierarchical Active Inference: A Theory of Motivated Control. *Trends in Cognitive Sciences*, 22(4), 294–306. <https://doi.org/10.1016/j.tics.2018.01.009>
- Proust, J. (2010). Metacognition. *Compass, Philosophy*, 11, 989–998. <https://doi.org/10.1111/j.1747-9991.2010.00340.x>

- Rademaker, R. L., Pearson, J., Kosslyn, S. M., & Kosslyn, S. M. (2012). Training visual imagery : improvements of metacognition , but not imagery strength. *Frontiers in Psychology*, 3(July), 1–11. <https://doi.org/10.3389/fpsyg.2012.00224>
- Rahnev, D., Adler, W. T., Akdogan, B., & Balci, F. (2019). *The confidence database*. August. <https://doi.org/10.31234/osf.io/h8tju>
- Rahnev, D., Balsdon, T., Charles, L., Gardelle, V. De, Denison, R., Desender, K., Faivre, N., Filevich, E., Jehee, J., Rahnev, D., Balsdon, T., Charles, L., Gardelle, V. De, & Denison, R. (2021). *Consensus goals for the field of visual metacognition*.
- Rahnev, D., & Denison, R. N. (2018). Suboptimality in Perceptual Decision Making. *Behavioral and Brain Sciences*, 1–107. <https://doi.org/10.1017/S0140525X18000936>
- Rahnev, D., & Fleming, S. M. (2019). How experimental procedures influence estimates of metacognitive ability. *Neuroscience of Consciousness*, 2019(1), 1–9. <https://doi.org/10.1093/nc/niz009>
- Rahnev, D., Koizumi, A., & Mccurdy, L. Y. (2015). Confidence Leak in Perceptual Decision Making. *Psychological Science*, October. <https://doi.org/10.1177/0956797615595037>
- Rausch, M., Hellmann, S., & Zehetleitner, M. (2018). Confidence in masked orientation judgments is informed by both evidence and visibility. *Atten Percept Psychophys*, 134–154.
- Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information : Effects on monitoring and control. *Psychonomic Bulletin & Review*, 16(3), 550–554. <https://doi.org/10.3758/PBR.16.3.550>
- Roebbers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, 45(May), 31–51. <https://doi.org/10.1016/j.dr.2017.04.001>
- Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Current Biology*, 28(24), 4014-4021.e8. <https://doi.org/10.1016/j.cub.2018.10.053>
- Rosenthal, D. (2005). *Consciousness and Mind*. Oxford University Press UK.
- Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, 10(1), 1–11. <https://doi.org/10.1038/s41467-019-09075-3>
- Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*, 1(May). <https://doi.org/10.1017/pen.2018.16>
- Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task

- Performance. *Biological Psychiatry*, 1–9.  
<https://doi.org/10.1016/j.biopsych.2017.12.017>
- Salem-garcia, N., Palminteri, S., & Lebreton, M. (2021). The computational origins of confidence biases in reinforcement learning. *PsyArXiv Preprints*.  
<https://doi.org/10.31234/osf.io/k7w38>
- Samaha, J., & Denison, R. (2022). The positive evidence bias in perceptual confidence is not post- decisional. *Neurosci Conscious*, 1(Jul).
- Schulz, L., Fleming, S. M., Dayan, P., Schulz, L., & Fleming, S. M. (2021). Metacognitive Computations for Information Search : Confidence in Control. *BioRxiv*, 1–35.  
<https://doi.org/10.1101/2021.03.01.433342>
- Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How Local and Global Metacognition Shape Mental Health. *Biological Psychiatry*, 18, 1–11.  
<https://doi.org/10.1016/j.biopsych.2021.05.013>
- Sepulveda, P., Usher, M., Davies, N., Benson, A., Ortoleva, P., & Martino, B. De. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *BioRxiv*, 2020.04.14.031971. <https://doi.org/10.1101/2020.04.14.031971>
- Seth, A. K., Dienes, Z., Cleeremans, A., Overgaard, M., & Pessoa, L. (2008). Measuring consciousness: relating behavioural and neurophysiological approaches. *Trends in Cognitive Sciences*, 12(8), 314–321. <https://doi.org/10.1016/j.tics.2008.04.008>
- Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, 18(4), 186–193.  
<https://doi.org/10.4324/9781315630502>
- Shekhar, M., & Rahnev, D. (2020). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, 1–12. <https://doi.org/10.1016/j.tics.2020.10.007>
- Sherman, M. T., Seth, A. K., Barrett, A. B., & Kanai, R. (2015). *Prior expectations facilitate metacognition for perceptual decision*. 35, 53–65.
- Smith, B. C. (2014). What does metacognition do for us? *Philosophy and Phenomenological Research*, 89(3), 727–735. <https://doi.org/10.1111/phpr.12148>
- Smith, J. D., Couchman, J. J., & Beran, M. J. (2012). The highs and lows of theoretical interpretation in animal-metacognition research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367, 1297–1309.  
<https://doi.org/10.1098/rstb.2011.0366>
- Smith, J. D., Redford, J. S., Beran, M. J., & Washburn, D. A. (2006). Dissociating uncertainty responses and reinforcement signals in the comparative study of uncertainty monitoring. *Journal of Experimental Psychology: General*, 135(2), 282–297.  
<https://doi.org/10.1037/0096-3445.135.2.282>

- Soutschek, A., Moisa, M., Ruff, C. C., & Tobler, P. N. (2021). Frontopolar theta oscillations link metacognition with prospective decision making. *Nature Communications*, 1–8. <https://doi.org/10.1038/s41467-021-24197-3>
- Spence, M. L., Dux, P. E., & Arnold, D. H. (2016). Computations underlying confidence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, 42(5), 671–682. <https://doi.org/10.1037/xhp0000179>
- Thompson, V. A. (2009). Dual -process theories : A metacognitive perspective. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond*.
- Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: Consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1412–1423. <https://doi.org/10.1098/rstb.2011.0421>
- Tomasello, M. (2022). *The Evolution of Agency: Behavioral Organization from Lizards to Humans*. The MIT Press,.
- Van Der Plas, E., David, A. S., & Fleming, S. M. (2019). Advice-taking as a bridge between decision neuroscience and mental capacity. *International Journal of Law and Psychiatry*, 3(June).
- Vlassova, A., Donkin, C., & Pearson, J. (2014). Unconscious information changes decision accuracy but not confidence. *PNAS*, 111(45), 16214–16218. <https://doi.org/10.1073/pnas.1403619111>
- Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., Dolan, R. J., & Blakemore, S. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, 22(1), 264–271. <https://doi.org/10.1016/j.concog.2013.01.004>
- Woelfle, M., Olliaro, P., & Todd, M. H. (2011). Open science is a research accelerator. *Nature Chemistry*, 3(10), 745–748. <https://doi.org/10.1038/nchem.1149>
- Yon, D. (2020). Enhanced metacognition for unexpected action outcomes. *PsyArXiv Preprints*, 1–10. <https://psyarxiv.com/3vn96>

# Bridge:

## From landscape to learning.

This thesis argues for the role of subjective value in providing a comprehensive picture of procedural metacognition as a thermostat for decision coherence. By developing the models of both the function and the computation of metacognitive monitoring signals, we suggest that subjective value is essential to close the loop and provide a comprehensive understanding of metacognition. The first half of the thesis presents a conceptual framework for the function of metacognitive monitoring signals.

Chapter 2 aimed at defining the concept of procedural metacognition across the fields of philosophy and cognitive neuroscience. Following the first chapter which offered a functional introduction of metacognition, this second chapter aimed at deepening this definition of procedural metacognition through both fields. More specifically, trying to detach from the traditional approach in cognitive neuroscience which defines metacognitive monitoring mainly by the computation of its signals based on their input, here, we used a multi-disciplinary approach to turn around the research perspective and looked at how these monitoring signals might serve different metacognitive functions.

First, relying on philosophical literature, we defined some of the unique controls and abilities that metacognition might bring to healthy human adults to function coherently both with themselves and with a society. We reviewed a three dimensional landscape where metacognitive monitoring signals are shaped by the (1) control function they serve together with the (2) representational and (3) conscious levels they require the agent to perform in order to compute them.

Secondly, we reduced the dimensionality of this space into a more linear one to discuss this procedural approach of metacognition in the cognitive-neuroscience literature. More specifically, looking at metacognition from both the computation of its input and its executive functions we reviewed the literature in the aim of sculpting

the boundaries of the different metacognitive signals in this multi-dimensional space. Thereby, in this chapter, we attempt at articulating the metacognitive space by defining both the function and computation of monitoring signals.

From this functional map of procedural metacognition, two central questions can be noted: In contexts where no action is required and no executive control has to be exerted to regulate behaviour, then does metacognitive monitoring still serve any function at all? And if some metacognitive monitoring occurs for another purpose than behavioural control, then does value-base input still serve a central role towards this thermostat-like function? To answer both these questions, we propose in Chapter 3 a review of the literature on both perceptual and value-based metacognition in regards to their contribution to inference. We propose a Bayesian architecture for the inference process where metacognition appears as an advantageous add on to tune learning by accounting for different sources of evidence reliability.

## Chapter 3

# Inferential Metacognition of Perceptual and Value-based Decisions.

### **Lexicon:**

**Global:** relating to a stationary context's causal structure which defines the local states. In a controlled laboratory setting, a causal structure can be represented with autocorrelated trials (e.g. 98% the salt is in more-dots shaker) or can be neutral with independent (and random) trials (e.g. 50% the more-dots shaker is on the right side). In value based tasks, payoff structures are the causal structures that define the outcome of each alternative item, in perceptual tasks, dispositional structures are the causal structures that define the state in which the alternative items present themselves. The uncertainty of a causal structure can be probabilistic as for a lottery.

**Local:** relating to the present state of the world. Cognitive processes relating to local states are best studied in tasks with independent and randomised trials to cancel out global features of stationary contexts. In the world a local state is finite but can provide an uncertain signal due to its strength or noise.

**Policy:** decision-rule based on one's global belief *e.g.* the agents consistently choose more-dots shaker because she believes it to generally render the higher-value

outcome or she consistently chooses the item on the right because she believes that this is where the target item would generally be. Policies are learned as **prior beliefs** by sampling from or getting feedback from a stationary context.

**Goal:** defines the currency to be valued and maximised when making a decision (*e.g.* accuracy, points, coins, welfare, salt...).

**Strategy:** value-based policy which defines the target item as a result of associating one's goal together with one's prior belief about the context's payoff structure (*e.g.* goal: value= salt; prior belief : salt = more dots ; strategy: more dots = target).

**Value-based decision (VD):** the agent samples from her uncertain prior belief about the context payoff structure to define the item to be targeted as rendering higher value. Preferential tasks use an array of familiar items (known payoff structure was previously learned) and combine them in independent and random pairs (no stationary context) to study economic decision-making. Reinforcement learning (RL) present few unknown items to study how participants adapt their strategy to stationary contexts with autocorrelated trials where they learn the payoff structures through choices and feedback.

**Perceptual decision (PD):** the agent samples an uncertain perceptual signal from the local state of the world to identify the target item amongst two alternatives. PD generally present independent trials to cancel out the influence of a prior belief on the sampling from the local state.

**Hierarchical decision:** decision where both the global payoff structure and the signal from the local state of the world are uncertain. A hierarchical decision combines a VD to define the target item and a PC to identify it in the local state of the world.

**Local confidence (C or LC):** metacognition evaluates the reliability of a decision by sampling from its sources of evidence independently from the decision process. We argue that participants can distinguish two composite confidence levels in a hierarchical decision by monitoring independently the reliability of the chosen strategy (*i.e.* decision expected value sampled from learned prior) and of the perceptual identification (*i.e.* decision expected accuracy sampled from local perceptual evidence). Local confidence levels are best studied controlled laboratory setting with independent trials to cancel out global influences of a stationary context



(*i.e.* generally perceptual and preferential decision-making tasks). Confidence levels are conventionally explicitly reported by human decision makers before an eventual feedback would be provided. In a psychometric model with stable difficulty, confidence computation can be calculated from the posterior probability of a gaussian representation as the evidence beyond the decision criterion supporting the choice (e.g. difference in dot number or value).

**Global confidence (GC):** metacognition evaluates the reliability of a stationary context by sampling from its various sources of input (e.g. reliability of reward or of perceptual evidence).

**Bayesian learning:** the agent samples or acts in in the local world to infer its global causal structure as prior expectations. Feedback together with the prior expectation about a decision form the predicted error (correct or wrong decision) to update with a learning rate the prior in return.

## 1. Introduction

How does insight help us to learn better? Imagine you are invited over for dinner and want to add some salt to your dish. You see two seemingly identical shakers on the table, except for the number of holes on their cap. You expect the policy to have the salt in the shaker with more holes; and therefore reach for shaker on the right side which appears to have more dots on top. After putting it back on the table, you suddenly have a doubt: “ Wait, did I take the shaker with the most dots? Thinking about it I am not sure anymore which shaker should contain the salt!” But why would you have such feelings of awkwardness or certainty now that you acted? And what should you do about such feelings of doubt or confidence?

In this article, we see such ecological decisions as a hierarchical (here dual) decision process (Fig. 1, Sarafyazd & Jazayeri, 2019). First, in accordance with the **global context**, the agent defines the strategy to follow in order to reach her goal. This process defines the values of alternative actions and is therefore seen as the **value-based** side of the choice (*i.e.* target the more-holes shaker). Second, at every **local interaction** with this context, the agent thrives at **perceptually** identifying the state of the environment to accordingly decide which item to choose (*e.g.* the right or left shaker). As with two sides of a coin, certainty in both these decisions will influence the behaviour: being sure of the policy, we might confidently pour more of the shaker’s content into our dish (*i.e.* invest in your choice) and by having no doubt in having picked our target shaker, we might do so quickly without looking or thinking twice. In other words, by tracking different sources of possible errors, confidence levels inform us about both the reliability of our local identification but also about the reliability of the overall context and therefore enable us to adjust our behaviour accordingly.

Research in cognitive neuroscience currently aims at building a computational model of confidence levels which could contribute to the inferential framework. Whilst the most popular model of confidence relies on its a post-hoc correlation with the choices’ probability of being perceptually correct (Pouget et al., 2016), a recent review stressed the need to distinguish such **descriptive measures** from explanatory models of confidence (Shekhar & Rahnev, 2020). More precisely, the authors

discussed a list of known predictors of confidence levels, such as awareness or confidence in the previous trial, that systematically contribute to the computation of confidence levels independently from the decision-making input. In this same review, these parsimonious models deprived from such systematic contributors of confidence were flagged to be limited for the investigation of functional aspects of metacognition such as its domain generality. In empirical research, this debate on domain generality relies on using the same parsimonious metrics of confidence ability (probability of being correct) to infer whether the same monitoring system applies in various tasks. It is therefore flagged that the disagreeing results on this question could result from an incomplete picture of the metacognitive system and its computation. In other words, to understand whether metacognition has an ubiquitous access to decisions from different domains, the parsimonious model distinguishing accuracy from error might not take into account the overarching function of metacognition which this research essentially aims to capture. By highlighting the wide range of systematic contributors to confidence levels, the authors remind collaborators that each model should answer one question: when comparing whether a given parameter and condition affects the participants' ability to know whether their perceptual decision is correct, using the same model is justified – but when asking whether a single metacognitive system should monitor different sources of evidence in a similar way, parsimonious models might, even if they bring a seemingly conclusive response, only answer part of the question. Here we suggest that by reframing metacognition in a functional framework of inferential learning and thereby by closing the loop between monitoring and control, we account for a more holistic picture of the workings of confidence and intrinsically reshape questions such as domain generality.

In this review, we first define the implications of looking at metacognition from the inferential learning framework and the possible contributions of such research to the field of confidence. We synthesise the key elements of inferential learning and relate them to the metacognitive function to discuss the relevance of monitoring the reliability of these learning mechanisms. In a second part, we then discuss empirical research on metacognition by defining the contribution of local decision confidence

to the learning process. In parallel, we discuss how this contribution of metacognition to cognitive learning shapes in turn our view of the metacognitive system itself. Finally, we propose that metacognition embraces the learning machinery by embodying its structure and review the recent literature that demonstrates metacognitive learning at a global level.

## **2. Inferential metacognition**

From its emergence in the field of education and memory, metacognitive functions for monitoring and controlling cognitive processes have been tightly linked to learning. Here we discuss the theoretical framework that links confidence to inference and their respective contributions to each other.

### **2.1. A multi-dimensional monitoring**

To address the limits of the above described descriptive models of confidence, such as tracking the objective accuracy of trials, recent review have zoomed out to account for a broader mechanistic model of confidence computation. As a foundation in this enterprise, a review by Fleming and Daw (2017) discussed the empirical evidence for metacognitive monitoring to have its own access to inputs as a second order process rather than a mere extension of cognitive processes. In this review, the authors highlight that confidence levels appear to be systematically informed by partly independent streams of sensory input (e.g. unconscious evidence) while also having additional streams of input (e.g. motor evidence). The authors conclude that further than simply relying on perceptual evidence to evaluate the probability of choices to be objectively accurate, the metacognitive system is a second order system that uses its own sources of input to evaluate the reliability of a choice. Along these lines, a second review presented metacognitive monitoring as doing more than tracking “probability correct” by being multi-dimensional. Using a linguistic task where participants had to associate meaning to foreign pairs of antonyms, A. Koriat (1976) demonstrated that high confidence was predicted by the popularity of the answers rather than their accuracy. Therefore, A. Koriat concluded that in such tasks where the notion of accuracy is unknown, both responses’ consensuality and confidence appear to rely on similar cues. In his

review, A. Koriat (2015) then proposed that confidence, further than tracking objective accuracy relies on tasks specific cues which, as a decision rule, defines the evidence and representations on which the agent should rely to decide, and that this same decision rule is used at the monitoring level to evaluate one's choice. The Self Consistency Model (SCM) proposed that confidence levels measure **decisions' reliability** by quantifying the evidence for the choice given the task specific decision rule (e.g. linguistic symbolism as proxy of accuracy). In other words, it can be argued that instead of being an objective measure of decisions accuracy, SCM proposes a model where context, decision-making and monitoring can be connected by a decision rule which defines task relevant cues and is likely to be shared by individuals having a similar background experience. In this framework, the useful descriptive models of confidence as "probability correct" are therefore a special case of laboratory controlled perceptual tasks where the decision rule both relies on objective identification and is known by all participants. More recently, a review further supported this proposal of confidence levels as a measure of cognitive reliability by describing them as necessary for the Global Workspace theory describing the integration of different input to give rise to consciousness (Shea & Frith, 2019). In line with the above SCM, the present review on inferential metacognition reflects this pragmatic multi-dimensional structure where the individual tracks the reliability of his choices in respect of a context specific decision rule for the task at hand. As for these reviews which put into perspective the descriptive function of confidence and offer a richer and more accurate computational model of its workings, we address in this review how the framework of learning can contribute to our understanding of metacognition.

## 2.2. The structure of cognitive inference

Before reviewing the advantageous contribution of confidence to learning, we start by laying a broad canvas of the workings of learning. In cognitive science, a popular consensus sees the brain as a predictive machine (Hohwy, 2013). Often described by a Bayesian framework, these theories define how the biological brain, limited in its sampling and processing capacities, builds from **local** interactions with the world a **global** model of its hidden causal structure. In this framework, a passive agent can

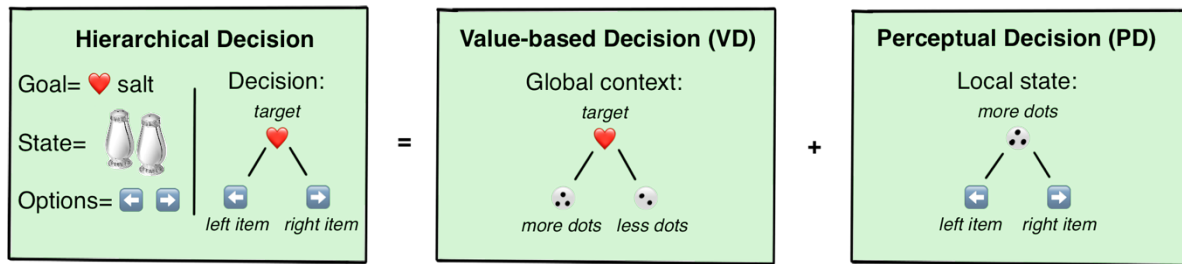
therefore for instance expect to see a face as convex (*i.e.* global prior for a **perceptual task**) and be tricked by an illusion when presented a concave sculpture of a face (*i.e.* local state of the world). This predictive framework is advantageous as it describes that the observer only needs to update her global expectations when they are at odd with the local world. This **prediction error** (or discrepancy between the global expectation and local outcome of the choice) is then used by the system to learn and update the global model, to in turn better fit the outside world. This automatic and generally unconscious update was demonstrated to be shared across several species and encoded at the neural level by dopamine levels both predicting outcomes and updating expectations accordingly.

By adding action as a powerful inferential tool to test one's causal model of the world, active inference describes how an agent also learns from the difference between the expected outcome of their actions and the actual local feedback they obtain. In this framework, the agent can therefore choose between options by comparing the expected outcome of different courses of action (*i.e.* **value-based task**). It was suggested that agents who take part in active inference only encode the causal links relevant to the agent's actions and goals as pragmatic strategies (Purcell & Kiani, 2016). In other words, if an observer is given the ability to act in the world, it is suggested that she will learn to fit strategies to given context based on relevant hidden states instead of encoding an objective holistic representation of the entire environment. As a result of this pragmatic framework, relevant actions' outcomes are typically evaluated for their utility in fulfilling the goal (*i.e.* value), and results a behavioural strategy. When studying confidence monitoring, all studies of metacognition are anchored in this framework where agents are required to compare and chose amongst options based on their expected value or accuracy. Also in line with Koriat's SCM (Koriat & Adiv, 2015) where decisions and confidence levels rely on a common rule, here we review the literature that suggests that metacognition monitors the decision reliability based on both the value-based and perceptual aspects. Similarly, in this inferential framework, we assume that decisions are hierarchical by relying both on perceptual evidence about the local state of the world and on internal evidence about the value of choice alternatives

(Sepulveda et al., 2020). The specific complementarity of these mechanisms, working as both sides of the decision-making coin, will be reviewed in more depth in part 2.1.

### **2.3. Monitoring reliability of different sources**

We discussed the multi-dimensionality of both metacognition and learning, which are closely intertwined in their contribution to behaviour and in their computational features. The obvious implication of metacognition in optimising learning is highlighted by the first publications in the field: they focussed on the educational setting and demonstrated the abilities of metacognition to both monitor and control knowledge. Besides this historical relevance of learning to metacognition, nowadays the popular theories describing the brain as a Bayesian predictive machine often discuss the role of metacognition as a central pillar of learning by advantageously monitoring the uncertainty in the system (Friston, 2011; Moulin & Souchay, 2015). As a measure of cognitive reliability, confidence can indeed be seen as an explicit monitoring of what predictive coding describes at implicit levels as the evidence's precision (Hohwy, 2012). This measure of reliability is essential in the Bayesian framework to combine information and therefore update the learned expectation at the global level. For instance, if in a dark restaurant you are only 60% sure you have grabbed the more-dots shaker you will certainly learn less from the outcome (it containing salt or pepper) than if you had all the time and perceptual evidence you wished and were sure 100% of the shaker you grabbed. Monitoring the reliabilities of both our perceptual identification and our decision rule can therefore contribute to adjusting the update of our beliefs based on such local feedback. The level at which this reliability monitoring needs to be conscious or explicit to contribute to learning is still poorly understood. Nonetheless it is transparent that metacognitive monitoring contributes to learning by enabling a weighting and manipulation of competing alternative expectations or beliefs. In the following part we discuss empirical work that describes how both perceptual and value-based confidence (PC and VC) contribute to refining learning.



**Figure 1: Complementarity of perceptual and value based decisions.** In our example, the agent finds herself in a foreign country (context) in which she wishes to take some salt, contained in either of two shakers: the *more-* or the *less-dots* shaker. This global rule is learned through experience by the agent who retrieves her knowledge when performing a value-based decision (VD). In the present state of the world, the shakers have equal chances to appear on either side of the table (flat prior), the agent samples local perceptual evidence to identify the target shaker as being on the left or right side (PD). The agent combines both complementary decisions to choose the valuable shaker.

### 3. Local confidence and learning

In this part, we review the evidence suggesting that confidence contributes to learning by refining the inference process and distinguishing between different sources of error.

#### 3.1. Complementarity monitoring of PC and VC

Before discussing literature focussing on the role of confidence to the learning mechanism (2.2), we discuss how even in tasks with independent trials where no prior expectations is to be built, empirical evidence supports the role of confidence as belonging to a larger scale inference system.

##### 3.1.1. Cognitive models of PD and VD

Just as black boxes, cognitive processes are best understood with clearly defined inputs and outputs. In the field of psychometrics, sequential sampling models (SSM) sheds light onto these cognitive processes: by assuming that the agent gradually accumulates evidence about the options up to a decision bound, these models capture various cognitive parameters (Ratcliff et al., 2016). Indeed, beyond input



parameters such as stimulus strength and noise, SSM fit for each participant (and decision) cognitive parameters such as the sensitivity of the subject to the evidence, her carefulness in choosing or how much delay she initially has in her decision process. These dynamic cognitive models were furthermore demonstrated to capture metacognitive factors that define confidence levels (De Martino et al., 2013).

To study decisions relying on subjective preferences instead of perceptual evidence, the field of economics uses familiar items, such as snacks, as stimuli from which the value and the uncertainty in value can be elicited from participants. In this field of research, a preferential decision is qualified as rational or coherent if the agent chooses the item which she rated with the highest liking. The operationalisation of two alternative forced choices with measurable items' values and uncertainties (relatable to evidence strength and noise in psychometrics) has enabled the application of SSM models to economic decisions. While being described by similar cognitive models, both PD and VD differ by the source of their input: anchored in perceptual evidence for PD and in memory and emotional cues in VD tasks (for reviews on studies of PD and VD see (Dutilh & Rieskamp, 2016; Shadlen & Shohamy, 2016; Summerfield & Tsetsos, 2012)).

In these preferential tasks, the decision rule is assumed to be stable together with the order of preference amongst all items during the experiment. This familiarity with items enable participants to retrieve from memory evidence in support of the decision rule for each item. This evidence accumulation process is therefore modelled similarly as in PD. In this review, we however will focus (from part 2.2) on dynamic setting with reinforcement learning (RL) tasks. In these tasks, instead of being presented with an extensive list of familiar items, participants are faced with a reduced list of items which value they must learn to maximize their reward by the end of the experiment. In both preferential and RL tasks, value evidence is therefore retrieved from global prior of expected value. On the contrary, in most PD tasks with independent and random trials where no prior expectation is to be built, the decision process is controlled to only essentially accumulate evidence from the local state of the world. As illustrated in Fig. 2, both VC and PD can therefore be seen as

complementary parallel processes that accumulate evidence from global learned priors for VD and from local input from the world state for PD (assumed flat perceptual priors). In RL tasks, switch between contexts enable to study the participants sensitivity of these higher-order strategy decision by measuring the fit of their behaviour with the changes in reward contingencies in a similar manner as would be studied behavioural sensitivity to change in world state in PD tasks (Fig 1).

Lastly, beyond borrowing a cognitive model to the field of psychometrics, the field of economics has itself developed SSM further to account for cognitive processes that are common to VD tasks. The Gaze-weighted Linear Accumulator Model (GLAM) takes in account the independent accumulation of evidence for both items based on the attention they receive by the participants, therefore accounting for the fact that the valuable (and often more salient) item is often most looked at and selected. A recent study by Sepulveda *et. al.* (2020) demonstrated that both PD and VD were best explained by attention-guided evidence accumulation. Therefore, at least for simplified tasks (Fig. 1 and 2), we assume that both PD and VD mechanisms are similar and parallel cognitive processes with complementary sources of input: based on the local state of the world for PD and based on globally memorised value for VD tasks.

### **3.1.2. Monitoring reliability: PC and VC**

In both PD and VD, confidence levels reflect an array of cues relating to the decision's reliability given the task's goal. As a central pillar to the computation of confidence levels, both PC and VC levels reflect the decision's difficulty as predicted by the stimuli. In PD, confidence reflects stimuli parameters such evidence strength (e.g. difference in dot number between both shakers) and noise (e.g. visibility or contrast of dots) (Bang & Fleming, 2018; Hebart et al., 2016). In VD, confidence levels also capture evidence strength as the difference in the items value (De Martino et al., 2013; Koriat, 2013), but, unlike in PD, it is debated whether value uncertainty contributes to confidence judgments (Brus et al., 2021; Castanheira et al., 2021; Quandt et al., 2021).

Beyond these Bayesian parameters of the stimuli, confidence also monitors other cues of the decision reliability relating to the stimuli. While reports of confidence in both PD and VD ask participants about the relative optimality “did you chose the accurate/ preferred/best item?”, confidence reports also capture cues about the overall relevance of the decision to the decision rule. In tasks with changing goal, Sepulveda *et. al.* (2020) found that confidence increases with the average value of items (VD) or of dot number (PD) when asked to select respectively favourite item or more-dot stimuli. However, this correlation was found to be inverted for opposite goals: selecting least favourite item or less-dot stimulus. While this average magnitude (*i.e.* value in VD or dots in PD) of both items does not matter to make the decisions, at the metacognitive level, confidence did encode the average relevance of the items set to the task goal (see also Lebreton *et al.*, 2019). Another non-normative cues of confidence were found in a two armed bandit task where the variance an item’s value only predicted confidence if it could bring a suboptimal choice closer to optimality as to make it relevant to the task’ goal (Hertz *et al.*, 2018). Other studies have also demonstrated that metacognition has privileged access to other streams of stimuli evidence (beyond what the decision process computes) in order to evaluate the decision reliability for the goal. Following the second-order model of confidence (Fleming & Daw, 2017), these additional streams concern for instance subliminal (Charles *et al.*, 2013; Cortese *et al.*, 2019) and post-decision evidence (Moran *et al.*, 2015; Navajas *et al.*, 2016).

Further than the items’ perceptual of value-based evidence, local confidence levels also reflect the reliability of the cognitive process at play in the decision-making process. More specifically, a recent preferential study suggested that VC was more influenced by noise in the cognitive process than by the uncertainty of items value. While in PD tasks the stimuli noise is often taken as proxy for confidence (see part 2.2.), this VD task demonstrated that instead VC was rather affected by attention to retrieve items value from memory and cognitive effort to compare these familiar items (Brus *et al.*, 2021). In a framework of perceptual and value-based evidence complementarity, it was suggested that the tight relation between VC ad attention could have a function of regulating the effort in accumulating perceptual evidence such as to adjust the level of effort to achieve the goal given the contextual difficulty

(Sepulveda et al., 2020). Another dimension of cognitive reliability is often captured by the heuristic of motor action or response time as a proxy for the fluency or sense of ease with which the decision is made (Charles et al., 2020; Fleming, 2016; Fleming et al., 2015, 2016; Kiani et al., 2014; Patel et al., 2012).

Together, these studies (with independent trials and no learning at play) highlight how both PC and VC evaluate the reliability of the decision for the goal rather than its normative optimality (correct-error or best-worst). In both PD and VD, but eventually with more weight in VC (Brus et al., 2021), metacognitive evaluation also takes into account the reliability of the cognitive process at play when making the decision. This monitoring of the cognitive process reliability can be argued to be pragmatically key: if the metacognitive function is ultimately to control cognition and behaviour to ensure that the goal is met, then monitoring whether the cognitive effort is up to the job is necessary to tune it in turn. These effects of local confidence on cognitive and behavioural control will be more extensively discussed in part 2.3.

### **3.1.3. A common currency of reliability**

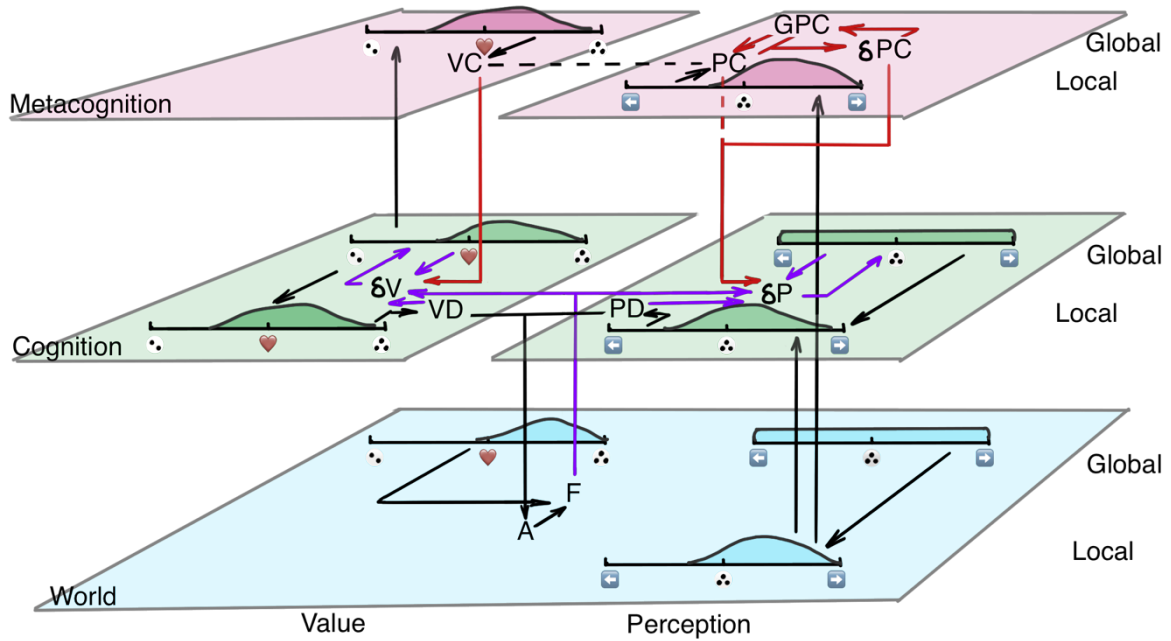
If PC and VC might monitor two complementary sides of a decision's reliability, then the question emerges about whether both signals are commensurable and could be integrated. While participants are sometimes incentivised to perform well at perceptual tasks (*i.e.* rewarding accuracy), the link between how participants evaluate their PD accuracy and value them was only recently studied empirically. In 2018 Lebreton et al. demonstrated that, although participants were asked to report the accuracy of their PD with their PC, participants also cued their PC on the magnitude of the incentives. The authors concluded that metacognitive monitoring tracks the choice value together with its perceptual accuracy as "two sides of the same coin". While this research highlights the link between a choice's PC and value, the distinctions and relations between local PC and VD remains is still poorly understood (see part 3 for studies in RL tasks). Indeed, for some real life decisions, the integration of perceptual and value cues might be done at very early stages through learned associations and heuristic and lead to a common evaluation of both perceptual and value reliabilities together.

This question of evidence integration according to their reliabilities is central to Bayesian inference and also investigated in the field of multi-sensory integration which could give an int on the metacognitive access to the reliability of different streams of evidence. Indeed while we discussed above the fact that metacognitive process can often access more evidence than the decision process itself (such as subliminal evidence), the question remains about whether confidence can access uncertainty such as whether two sources of evidence should or not be integrated together (Deroy et al., 2016; Dijkstra et al., 2022). Such ability dissect the reliability of distinct streams of perceptual evidence could provide additional evidence about the reliability of our choices and how to learn from our environment. Fairhurst et al. (2018) put to the test whether participants' confidence in touch and vision was affected similarly by an illusory T shape. The authors demonstrated that participants confidence in both modalities changed with the strength of the illusion: participants were more confident in their visual than tactile decisions when the illusion was weak, but more confident in their tactile than visual decisions when the illusion was strong. These results suggest that, depending on the contexts, participants rely differently on their sensory modalities based on how reliable they judge it to be. On the contrary, another illusion by Travers et al. (2020) was made to study the racial bias: the illusion of seeing darker greys on a face with African features compared to a Caucasian face, despite a controlled greyscale. These heuristic affecting early evidence integration were studied in other experiments to demonstrate how the racial bias associates threat with Afro American individuals (expected to be more likely to hold a gun than Caucasian individuals, Amodio & Devine, 2006; Plant & Peruche, 2005). Similarly, the question remains of whether such heuristics could link perception to some expected value and whether metacognition could access the reliability of perceptual and value evidence independently.

Together with this question of confidence commensurability and integration comes the question of domain generality. In perceptual tasks, the question is raised of whether, despite the same metacognitive computation or cerebral system would compute confidence levels for decisions relying on different sensory inputs. While

the computational question can be investigated with behavioural experiments, the investigation about its cerebral substrates requires a neuro-imaging approach (Rouault, McWilliams, et al., 2018). In the previous part we discussed the computational evidence comparing PC and VC in local decision which demonstrated both similarities but also distinctions such as in their functions in behaviour (see part 2.3 for more). Regarding the cerebral substrates computing PC and VC, the existing literature also supports some similarities, such as in the involvement of the rostral-lateral prefrontal cortex (rlPFC) used for explicit reports, and some distinctions such as the recruitment of the anterior cingulate cortex (ACC) in for PC and of the ventro-medial prefrontal cortex (vmPFC) for VC (De Martino et al., 2013; Fleming & Dolan, 2012).

The studies discussed so far (part 2.1.1-3) focussed on comparable experiments in which the computations of PC and VC can be modelled similarly (independent trials of local decision with no learning). While the role of confidence in these tasks will be discussed for their role in cognitive and behavioural control in part 2.3, we will now focus on the RL literature where confidence can be studied for its contribution to learning in hierarchical tasks (Fig 1). In the following literature, we focus on simple learning tasks where perception strives to identify the local state of the world (with no prior to be learned) and inference evaluates the contextual reward contingency to target the best item for the goal (Sarafyazd & Jazayeri, 2019). The state of the literature so far with no learning is simplified and illustrated in Fig. 2 with black computations. The following parts will discuss the learning literature (Fig 2. purple computation) and the contributions of confidence to this inference process (Fig 2. red computations).



**Figure 2: Schematic computation of the interactions between metacognition and inference.** Three levels (world in blue, cognition in green and metacognition in magenta) embody the causal structure of the world: the global contextual policies dictate (with noise) the local moment-to-moment states of the world. Cognition uses two complementary evaluative processes to make decisions: it samples the world’s uncertain perceptual signal to identify its present state and it retrieves an uncertain prior belief about the value of both items to choose the one to target (*i.e.* strategy). In the illustrated example, the agent’s goal is to obtain salt. His prior belief is to find this valuable salt in the more-dots shaker. When making a local value-based decision ( $VD$ ), the agent retrieves this learned value of both shakers from memory and decides to target the more-dots shaker. In parallel, sampling from the local state of the world’s signal (with no prior belief about finding it on either of both sides), noisy evidence is accumulated and the identification results in a perceptual decision ( $PD$ ) for the item on the right side. From the combination of both  $PD$  and  $VD$ , (respectively about the local state and global reward contingency) the agent reports her decision through an action ( $A$ ). At the metacognitive level, evidence is accumulated to evaluate the reliability of both  $PD$  and  $VD$  and to form confidence reports. Coming from their own stream of evidence,  $VC$  is built on the uncertain prior belief that supports targeting the more-dots shaker for salt, and  $PC$  is built on the uncertain local signal that supports that this shaker is on the right side. Both

confidence levels can eventually leak into each other or be integrated depending on the task at hand. In learning tasks (purple computations) the action resulting in feedback (F) can be compared with the prior expectation given the decision to form a prediction error  $\delta$ . The agent can learn from this latter by updating his global priors (with rate  $\alpha$ ). The interactions between inference and metacognition are illustrated by red computations. Local confidence levels (or their estimated proxy from stimuli features) can be used to refine learning such as by identifying whether an error might result from an inaccurate perceptual identification or an inaccurate belief about the current reward policy (proposed computation as by Sarafyazd & Jazayeri (2019):  $\alpha * \delta * VC * PC$ ). At the global level, metacognition can mirror the hierarchical structure of cognitive learning by extracting global confidence levels (GPC) from local confidence levels PC. The confidence prediction error  $\delta C$  represent unexpected errors in a context: it is the difference between the expected reliability of decisions in the context (CPG) and the monitored reliability of the local decision (PC) (Guggenmos et al., 2016). This same  $\delta C$  is used to update GPC from local confidence levels in an inference computation (learning rate  $\alpha C$ ) that mirrors the architecture of cognitive inference. Lastly, GPC leaks into local PC to tune it to the contextual reliability of decisions.

### 3.2. PC and VC in learning

While older learning models (such as win-stay-lose-switch) predict behaviour in simple tasks, the development of the POMDP model (Partially Observable Markov Decision Process) captures best behaviour in more complex tasks with perceptual noise (Lak et al., 2017). Based on Bayesian framework, this model demonstrated that two monkey accounted in their learning (traditionally left side of Fig.2) for their belief state of the local perceptual representation (right side of Fig. 2). Indeed, while in older models of learning were based on prediction error as a binary difference between the expected and the obtained reward (win-lose), the present model uses perceptual ambiguity to refine the expected reward on a more continuous scale. POMDP model therefore borrows psychometric methods of uncertainty weighting to refine reinforcement learning models. However, the authors focussing on the neurobiology of this uncertainty encoding in animal models approximated the



concept of confidence to the stimuli perceptual noise. Going against the previously discussed definition of confidence as a reflective process about the decision made (Fleming & Daw, 2017; Pouget et al., 2016), the authors argue for this proxy to be a statistical definition of confidence since it predicts the monkeys probability to be correct in their choices. In this model nonetheless, the addition of this measure of PC as a factor of expected value was repetitively demonstrated to advantageously predict the monkeys learning and behaviour. In other words, the model explains that monkeys learned from the outcome of their decisions proportionally to their confidence level in their perceptual choices. In an ambiguous trials where the correct responses were unsure, receiving a reward was therefore less informative about the context contingency than when the perceptual evidence supported fully the decision.

Another study on monkey investigated the effect of further modelling the effect of this perceptual uncertainty but over the course successive trials in predicting the behavioural switch between strategies (Sarafyazd & Jazayeri, 2019). For this hierarchical task with both a local perceptual decision and a global strategy decision (Fig 1), monkeys were asked at each trials to report which of 2 rules they believed they had to follow (VD) before being presented a visual stimulus and reporting accordingly their local decision (PD given VD). In this task, it was observed that the probability to switch strategy increased after successive negative feedback: monkeys monitor the reliability of their VD and when it becomes too low they decide to switch to the alternative strategy. One could suggest that by monitoring in parallel the expected reliabilities of both VD and PD in the present trial, a comparative computation could also enable to infer the source of errors and tune learning even more finely. In the material and methods, the authors define how hierarchical models of decision-making can be implemented to take into account the confidence about the strategy VC together with the perceptual confidence PC. An ideal observer model is first described as taking both PC and VC as factors of expected value to update the belief about the decision rule. However, the lack of data in the present study constrain the authors to fitting a simpler model where confidence in decision rule VC decreases linearly with the number of errors and is reset to null after a

positive feedback. In Figure 2, we suggest the authors' first ideal observer model where both VC and PC can interact together, and also interact with learning at the cognitive level. Testing this model would therefore be central to understanding the interaction between confidence and learning and do so even further if done in humans with the more common definition of confidence as verbal reports from a reflective processes on decisions.

### **3.3. Local control**

The monitoring of the decisions reliability regarding their perception, learned values and cognitive processes makes confidence levels some powerful cues to improve the reliability of future decisions. First we discuss the optimisation of the cognitive reliability. In multiple studies, confidence was demonstrated to control the speed-accuracy trade off which defines the amount of evidence the individual accumulates before taking a decision (Baldson et al., 2020; Desender et al., 2019). One such example is captured by the modelling of an increase in required evidence after low confidence report, which in turn guided post-decision evidence accumulation to either stick to or change one's mind about the PD (Desender et al., 2020; Schulz et al., 2021). Beyond the control of such cognitive process, metacognition was also demonstrated to be able to learn completely independently from the evidence which the cognitive level has access to. A brilliant experiment demonstrated the capacity of metacognition to monitor the reliability of PD despite the fact that cognition was limited by its lack of access to subliminal evidence (Cortese et al., 2020). While participants could not learn to perform better than chance level, they nonetheless learned to attribute wages appropriately given this subliminal cues to increase their income over time. In other words, by having a refined access to the reliability of the decision making process, metacognition could still skilfully provide high confidence in correct decisions. Together, these both lines of study suggest that metacognition can both guide cognitive learning by tuning the cognitive processes at play, and also learn independently from it by accessing more complex evidence. Another line of research concerns the question of cognitive control in the decision. While the links between sense of agency and metacognitive monitoring are not empirically tested to our knowledge, an interesting study proposed to look at the link between inserted

errors and post decision slow-down (Logan & Crump, 2010). The study demonstrated typists could detect when they made an error and did not account similarly when a random error was inserted. This result suggest that participants can infer about how skilled they are at a task independently from other environmental factors determining the outcome of the action. This ability to monitor cognitive performance independently from its outcome is also important to know whether one should strive to improve cognitive ability or not.

Secondly local confidence in such independent trials was also demonstrated to adjust future behaviour and decisions. on the behavioural side, local confidence was also demonstrated to predict change of mind and transitivity optimisation in independent trials (Erik & Folke, 2017; Resulaj et al., 2009). While these trials-to-trials adjustments rely on the sensitivity of confidence to accurately estimate the reliability of a decision, it was demonstrated that psychological disorders are associated with mis calibrated confidence bias rather than sensitivity (Seow et al., 2021). In the following part, we will address global confidence levels in learning and discuss how these contextual cues, observable as a confidence bias contribute to learning and behavioural control in healthy participants or relate to clinical spectrums.

#### **4. Global confidence**

##### **4.1. Monitoring global confidence**

As the cognitive structure mirrors the world's causal structure by differentiating local states from global priors, metacognition also embodies the hierarchical structure of the world by building global estimates of reliability: global confidence levels. After we discussed how local levels of confidence contribute to cognitive learning, here we synthetise the literature revealing a global monitoring of reliability. In other words, these tasks require participants to infer, further than the strategy to follow, the reliability or probability of the strategy to lead to success.

#### **4.1.1. Monitoring expected sources of errors**

In perceptual tasks, Rouault et al. (2019) randomly presented participants with trials from 2 tasks identified by different colour framing. In both tasks participants had to choose the stimuli with more dots and report their local confidence in their local choices but also had to report at the end of the block the task on which they wished to be rewarded for their performance. In one condition, participants received feedback on every few trials and successfully manage to estimate their average performance on each of both colour coded task and chose the one at which they performed best to be rewarded on. The interesting finding is that in another condition of the experiment, instead of receiving feedback every few trials, participants had to report their local confidence on the perceptual choice they just made. In this condition too, participants successfully managed to report a lower average performance at the hard task and a higher average performance on the easy task which they selected for reward. While data were not sufficient to draw conclusive models (see part 3.1.2), the authors suggest that the global confidence levels for the tasks were informed by the local reports of confidence from trials to trials. More recently, this result was replicated by also demonstrating that the more trials the participants had in a task, the better they were at estimating their global performance at the task. This later results therefore highlight the importance of local confidence in informing global confidence estimates (Lee et al., 2021).

In value-based task on the other side, we argue that confidence (VC) is inherently global. Value based decisions tasks can be divided in two types: preferential – with uncorrelated trials where the value depends on the participants own past experience- and reinforcement learning (RL) tasks where participants are presented repeatedly with new items in order to learn their values during the experiment. As discussed before, the monitoring of this value side of the “ two-sided decision-making coin” therefore aims at investigating the fit of the agent’s global knowledge with the actual contextual reward policies, whereas the perceptual side informs about the fit with the local perceived state. Therefore, while GPC can be seen to learn from cumulated local confidence in perceptual tasks with independent trials, RL tasks make VC global in nature by relying on the accumulation of evidence over trials

at the cognitive level (Fig. 2). Indeed by investigating VD through RL tasks, VC appears to be shaped by the global evidence accumulated during the learning phase. Both preferential and RL tasks, the evidence on which VD relies is sampled from various past experience and sampling.

It remains to be studied whether in both preferential and RL tasks, VC presents similar computations and relation to behavioural control. Indeed while in RL tasks rewards can be neatly controlled by the experimenter, it is known that for instance the participant's socio-economical status can affect the subjective representation of the items' value. While socio-economical status might also shape how value is learned for familiar items of preferential tasks, the fact that values here are reported by the participants rather than learned from controlled stimuli might be one of the many cues that can create difference between both approaches of VD tasks. Furthermore, empirical results disagree on the effect of value uncertainty on decision confidence between both types of VD tasks. While in preferential tasks, empirical evidence is inconclusive about whether variance in value elicitation affects VC (Lebreton et al., 2015; Polanía et al., 2019) or not (Brus et al., 2021; De Martino et al., 2013) in RL tasks, the variance of items reward was shown to be captured by decision confidence and to guide exploration / exploitation behaviours based on these levels of confidence (Boldt et al., 2017; Hertz et al., 2018). Therefore, to better understand the role of confidence in learning and behavioural control in naturalistic environment, it would be very useful to investigate and bridge further the operationalisation of concepts as uncertainty in preferential and in RL tasks. In this endeavour, Folke et al. (2017) selected for this experiment the familiar items based on each participant's reported subjective value. This method could be used further to compare the building and role of confidence between preferential and RL tasks. A central difference between preferential tasks, where trials are independent and more "local" (as in PD tasks), and RL tasks which are about higher-order strategy selection, regards the evidence on which VC relies. In preferential task, VC represent the reliability of the present choice in relation to its alternative, whereas in RL task VC this summary statistic (Pouget et al., 2016) is accumulated from evidence amongst several trials to track the reliability of the applied strategy. In both VD tasks, VC tracks the likelihood of the choice to be successful based value learned in a stable

context. Furthermore, whether trials are independent (i.e. preferential tasks) or forming a stable context within the experiment (i.e. RL tasks), VC monitors the reliability of decision as undeniably defined by the agent's goal (Castegnetti et al., 2021).

#### **4.1.2. Metacognitive inference: learning GC from LC**

How do agents manage to form global levels of confidence over time and track the reliability of a given source of error? The drift of perceptual tasks is have independent trials from which the participants should not be able to learn where to find the correct item. In a perceptual learning task however, it was demonstrated that participants managed to improve their ability to visually discriminate Gabor patches and thereby increase their performance (Guggenmos et al., 2016). In this task, a learning model demonstrated that this learning was best explained by a model where not only cognitive perception learned but the metacognitive level too by forming global levels of perceptual confidence (GPC). This mirrored computation between the cognitive and metacognitive levels contributed to learning on two points. First, as mentioned above, local PC monitored each trial's perceptual noise while, and from them, global GPC learned about the level of noise to be expected in this task. At each trial both local and global PC were compared together to produce a confidence prediction error ( $\delta PC$ ) being positive if the local PC was greater than predicted by GPC, and negative otherwise. Conceptually, this hierarchical metacognition (local – global) enables the agent to monitor whether the reliability of the present trial can be (or not) accounted for by the expected reliability in the present context. Importantly, this  $\delta PC$  was demonstrated to best explain behavioural learning at two levels (Fig 2): first by modulating the valance of the metacognitive learning rate from local to global confidence as to increase or decrease expected reliability; secondly, by modulating the magnitude of the perceptual learning as by increasing or decreasing the weight of an error according to this contextual expectation. Altogether this research therefore supports the fact that metacognition informs the inferential processes by mirroring the cognitive

Bayesian structure of updates and also by interacting directly with it. More recently, Rouault et al. also proposed hypothetical model of metacognitive inference, however these latter remain to be empirically tested (Rouault et al., 2019).

While we present a simple example of a two level hierarchical task (Fig 1 and 2) with local perception and global value, we argue that it does provides the dimensions of a wider monitoring machinery where even higher levels of hierarchy can apply. For instance, in our example, further than having a VC about the strategy to target the more-dot shaker, one could have an even more higher-order level of confidence such as about our own reliability at knowing what is good for our own diet; or the reliability of our moral compass when it comes to consuming certain food. Ultimately these higher and higher levels of reliability monitoring, both in value-based and perceptual tasks, form self-efficacy beliefs which can inform an agent about his ability to perform in different tasks and environments. The importance of these higher level monitoring on guiding behaviour and psychological health will be discussed in part 3.2.

#### **4.1.3. GC in turn tunes LC**

The computation of local confidence goes beyond tracking the reliability of the local decision. While in perceptual tasks (presenting independent trials), a leak from previous to current level of confidence was often described as a metacognitive inefficiency, Rahnev et al. (2016) highlights: “Specifically, it has been argued that observers assume the world is autocorrelated because it usually is“. Here we could suggest that instead of a trial to trial leak, the effect of previous trials could be mediated hierarchically by the GPC learning. This contextual cues from GPC such as repetitively predicting lower outcomes over few trials can signal a need to change strategy or task (Sarafyazd & Jazayeri, 2019). On the other hand, while such low reliability in a few trials does not exceed a certain threshold (e.g. expected reliability in the given tasks) this GC enables the agent to remain resilient and stick to the current strategy, thereby adapting behavioural flexibility to the reliability of contextual policies

Local confidence also relies on other cognitive cues which can serve as heuristic in defining the reliability of a choice and which can be dependent on a specific context. Mainly, response time which is widely acknowledged as a marker of cognitive effort is known to correlate with local confidence levels. It was suggested that this cue could serve as a supramodal common currency across various decisions making domains to infer about decision reliability (Faivre et al., 2018). Similarly, it could be argued that the lack of reliability in prior expectations (such as uncertainty in the strategy to follow: low VC), or the cues on which to base one's decision (Koriat, 1976; Koriat & Adiv, 2015), can increase the required time and effort to make a decision. While taking longer time on these harder decisions might result in a similar performance to faster decisions in easier contexts, the heuristic link between confidence and response time can be a marker that, at higher metacognitive level, the strategy reliability is low and bringing additional uncertainty in the cognitive process. In other words, cues regarded as heuristics of local perceptual confidence could be markers of higher order metacognitive uncertainty about the task such as uncertainty about the strategy to follow (i.e. low VC).

#### **4.2. Global control**

Learning tasks provide insight in higher-order action selection: global strategies. Accurate learning can be operationalised by evaluating whether an agent is able to adaptively switch strategy in tune with the contextual reward contingencies. Value based confidence was demonstrated to reflect the stochasticity in the value of items together with guiding the exploration-exploitation switch (Boldt et al., 2017). This higher order metacognitive monitoring is therefore argued to be tightly linked with behavioural control. While the role of GPC in task selection was demonstrated in multiple studies (Lee et al., 2021; Rouault et al., 2019), it would be interesting to study whether another type of GPC, namely confidence in learned perceptual priors, would have also an effect on behavioural control by for instance selecting the task with the most reliable priors (e.g. the task where the most-dot shaker is always on the left side).



Regarding clinical populations, it was found that OCD patients had the same metacognitive ability (VC monitoring reliability of context policy) as healthy participants. Nonetheless, these OCD patients appeared not to be able to use these cues to control their behaviour (Vaghi et al., 2017). As previously discussed for local PC, a large study on psychopathologic traits also demonstrated that metacognitive ability in these local PC was not a predictor of either anxiety-depression, compulsive behaviour or social withdrawal traits (Rouault, Seow, et al., 2018). On the contrary, this later study suggested that it was instead a bias in PC that predicted these traits: a negative PC bias was predictor of anxiety-depression and a positive bias predictive of compulsive behaviours. Such subjective markers in confidence reports were recently suggested to be linked with higher-order metacognitive monitoring: global monitoring (Seow et al., 2021). Indeed, these bias in local PC could be seen as markers of over or under confidence in the strategy to follow. For instance, depressed individuals' low reported PC could reflect their low VC in knowing which strategy to apply to make valuable decisions. On the contrary, compulsive individuals with high PC could represent their over-estimation of the reliability of the strategy on which they rely (positively biased VC). Along these lines, this review argues that when investigating behavioural psychopathies, studies should go beyond the local moment to moment cues on which confidence rely to capture higher-order global metacognitive functions such as self-efficiency beliefs that are likely to be a relevant factor of the observed disabilities. Furthermore, it is suggested that while it was demonstrated that local metacognitive monitoring could be trained and improved, metacognitive therapies which are becoming increasingly popular to help suboptimal behaviour, would rather benefit from a finer understanding and training of more global levels of metacognitive monitoring. In this review, we aimed at highlighting the recent research that is concerned with understanding the functional aspect of metacognition as a learning machine to refine inferential learning and behavioural control. We proposed a conceptual map of the findings so far in the field in the hope to motivate research on these testable hypotheses to ultimately support the development of the study of metacognition as a useful psychological lever to help behavioural control.

## **5. Conclusion**

This review brings together the study of confidence in different tasks: perceptual and value-based, and illustrate their complementary contribution to the metacognitive system as regards to its role in learning. Rooted in the view that the brain is a predictive machine, we take the view that metacognition is an advantageous tuning system to learn more efficiently and in turn tune one's behaviour to different sources of reliability in the environment. We suggest in a simplified framework that the laboratory controlled decision making tasks reflect two distinct sides of Bayesian update: perceptual confidence monitoring the reliability of local state identification, and value based confidence monitoring the reliability of global evidence as learned over time and retrieved from memory such as to guide the course of action with strategy selection. Lastly, we highlight recent promising work on this hierarchical view of metacognition as to better understand and eventually help clinical population.

## References

- Amodio, D. M., & Devine, P. G. (2006). Stereotyping and Evaluation in Implicit Race Bias : Evidence for Independent Constructs and Unique Effects on Behavior. *Journal of Personality and Social Psychology*, *91*(4), 652–661.  
<https://doi.org/10.1037/0022-3514.91.4.652>
- Balsdon, T., Wyart, V., & Mamassian, P. (2020). Confidence controls perceptual evidence accumulation. *Nature Communications*, *11*, 1–11.
- Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 6082–6087.  
<https://doi.org/10.1073/pnas.1800795115>
- Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8), 1520–1531.  
<https://doi.org/10.1037/xhp0000404>
- Brus, J., Aebersold, H., Grueschow, M., & Polania, R. (2021). Sources of confidence in value-based choice. *BioRxiv*, 1–22.
- Castanheira, S., Fleming, S. M., & Otto, A. R. (2021). Confidence in risky value-based choice. *The Psychonomic Society*.
- Castegnetti, G., Zurita, M., & Martino, B. De. (2021). How usefulness shapes neural representations during goal-directed behavior. *Science Advances*, *7*, 1–13.
- Charles, L., Chardin, C., & Haggard, P. (2020). Evidence for metacognitive bias in perception of voluntary action. *Cognition*, *194*.
- Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94. <https://doi.org/10.1016/j.neuroimage.2013.01.054>
- Cortese, A., Lau, H. C., & Kawato, M. (2019). Metacognition facilitates the exploitation of unconscious brain states. *BioRxiv*, 548941.  
<https://doi.org/10.1101/548941>
- Cortese, A., Lau, H., & Kawato, M. (2020). Unconscious reinforcement learning of hidden brain states supported by confidence. *Nature Communications*, *11*, 1–

14. <https://doi.org/10.1038/s41467-020-17828-8>
- De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.  
<https://doi.org/10.1038/nn.3279>
- Deroy, O., Spence, C., & Noppeney, U. (2016). Metacognition in Multisensory Perception. *Trends in Cognitive Sciences*, *20*(10), 736–747.  
<https://doi.org/10.1016/j.tics.2016.08.006>
- Desender, K., Boldt, A., Verguts, T., & Donner, T. H. (2019). Confidence predicts speed-accuracy tradeoff for subsequent decisions. *ELife*, *8*.  
<https://doi.org/10.7554/eLife.43499>
- Desender, K., Donner, T. H., & Verguts, T. (2020). Dynamic expressions of confidence within an evidence accumulation. *BioRxiv*, 1–31.  
<https://doi.org/10.1101/2020.02.18.953778>
- Dijkstra, N., Kok, P., & Fleming, S. M. (2022). Perceptual reality monitoring: Neural mechanisms dissociating imagination from reality. *Neuroscience & Biobehavioral Reviews*, *135*. <https://doi.org/10.1016/j.neubiorev.2022.104557>
- Dutilh, G., & Rieskamp, J. (2016). Comparing perceptual and preferential decision making. *23*, 723–737. <https://doi.org/10.3758/s13423-015-0941-1>
- Erik, N., & Folke, T. (2017). *The Pragmatics of Confidence in Perceptual and Value-based Choice*.
- Fairhurst, M. T., Travers, E., Hayward, V., & Deroy, O. (2018). Confidence is higher in touch than in vision in cases of perceptual ambiguity. *Scientific Reports*, *8*(1), 1–9. <https://doi.org/10.1038/s41598-018-34052-z>
- Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. (2018). Behavioural, modeling, and electrophysiological evidence for domain-generalty in human metacognition. *The Journal of Neuroscience*, *38*(2), 263–277.  
<https://doi.org/10.1523/JNEUROSCI.0322-17.2017>
- Fleming, S. M. (2016). *Changing our minds about*. 3–5.  
<https://doi.org/10.7554/eLife.11946>
- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. <https://doi.org/10.1037/rev0000045>

- Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594), 1338–1349. <https://doi.org/10.1098/rstb.2011.0417>
- Fleming, S. M., Maniscalco, B., Ko, Y., Amendi, N., Ro, T., & Lau, H. (2015). Action-Specific Disruption of Perceptual Confidence. *Psychological Science*, 26(1), 89–98. <https://doi.org/10.1177/0956797614557697>
- Fleming, S. M., Massoni, S., Gajdos, T., & Vergnaud, J. (2016). Metacognition about the past and future : quantifying common and distinct influences on prospective and retrospective judgments of self-performance. *Neuroscience of Consciousness*, 1–12. <https://doi.org/10.1093/nc/niw018>
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1), 17–19. <https://doi.org/10.1038/s41562-016-0002>
- Friston, K. (2011). Embodied Inference : or “ I think therefore I am , if I am what I think .” *The Implications of Embodiment (Cognition and Communication)*, January 2011, 89–125.
- Guggenmos, M., Wilbertz, G., Hebart, M. N., & Sterzer, P. (2016). Mesolimbic confidence signals guide perceptual learning in the absence of external feedback. *ELife*, 5(MARCH2016), 1–19. <https://doi.org/10.7554/eLife.13388>
- Hebart, M. N., Schriever, Y., Donner, T. H., & Haynes, J. D. (2016). The Relationship between Perceptual Decision Variables and Confidence in the Human Brain. *Cerebral Cortex*, 26(1), 118–130. <https://doi.org/10.1093/cercor/bhu181>
- Hertz, U., Bahrami, B., & Keramati, M. (2018). Stochastic satisficing account of confidence in uncertain value-based decisions. *PLoS ONE*, 13(4), 1–23. <https://doi.org/10.1371/journal.pone.0195399>
- Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, 3(APR), 1–14. <https://doi.org/10.3389/fpsyg.2012.00096>
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Kiani, R., Corthell, L., & Shadlen, M. N. (2014). Choice Certainty Is Informed by Both Evidence and Decision Time. *Neuron*, 84(6), 1329–1342. <https://doi.org/10.1016/j.neuron.2014.12.015>

- Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition*, 4(3), 244–248.  
<https://doi.org/10.3758/BF03213170>
- Koriat, A. (2013). Confidence in Personal Preferences. *Journal of Behavioral Decision Making*, 26(3), 247–259. <https://doi.org/10.1002/bdm.1758>
- Koriat, A., & Adiv, S. (2015). The self-consistency theory of subjective confidence. *Oxford Handbook of Metamemory*, 1(June), 1–25.  
<https://doi.org/10.1093/oxfordhb/9780199336746.013.18>
- Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Current Biology*, 27(6), 821–832.  
<https://doi.org/10.1016/j.cub.2017.02.026>
- Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, 18(8), 1159–1167. <https://doi.org/10.1038/nn.4064>
- Lebreton, M., Bacily, K., Palminteri, S., & Engelmann, J. B. (2019). Contextual influence on confidence judgments in human reinforcement learning. *PLoS Comput Biol*, 15(4), 1–27.
- Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., Holst, R. J. Van, & Luigjes, J. (2018). Two sides of the same coin : Monetary incentives concurrently improve and bias confidence judgments. *Science Advance*, May.
- Lee, A. L. F., de Gardelle, V., & Mamassian, P. (2021). Global visual confidence. *Psychonomic Bulletin and Review*, 28(4), 1233–1242.  
<https://doi.org/10.3758/s13423-020-01869-7>
- Logan, G. D., & Crump, M. J. C. (2010). Cognitive illusions of authorship reveal hierarchical error detection in skilled typists. *Science*, 330(6004), 683–686.  
<https://doi.org/10.1126/science.1190483>
- Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive Psychology*, 78, 99–147.  
<https://doi.org/10.1016/j.cogpsych.2015.01.002>

- Moulin, C., & Souchay, C. (2015). An active inference and epistemic value view of metacognition. *Cognitive Neuroscience*, 6(4), 221–222.  
<https://doi.org/10.1080/17588928.2015.1051015>
- Navajas, J., Bahrami, B., & Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences*, 11, 55–60.  
<https://doi.org/10.1016/j.cobeha.2016.05.005>
- Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748), 4853–4860. <https://doi.org/10.1098/rspb.2012.1847>
- Plant, E. A., & Peruche, B. M. (2005). The Consequences of Race for Police Officers' Responses to Criminal Suspects. *Psychological Science*, 16(3).  
<https://doi.org/10.1111/j.0956-7976.2005.00800.x>
- Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, 22(1), 134–142. <https://doi.org/10.1038/s41593-018-0292-0>
- Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3), 366–374. <https://doi.org/10.1038/nn.4240>
- Purcell, B. A., & Kiani, R. (2016). Hierarchical decision processes that operate over distinct timescales underlie choice and changes in strategy. *Proceedings of the National Academy of Sciences of the United States of America*, 113(31), E4531–E4540. <https://doi.org/10.1073/pnas.1524685113>
- Quandt, J., Holland, R. W., & Veling, H. (2021). Confidence in evaluations and value-based decisions reflects variation in experienced values. *American Psychological Association*, September. <https://doi.org/10.1037/xge0001102>
- Rahnev, D., Evan, D., Riddle, J., Sue, A., & Esposito, M. D. (2016). Causal evidence for frontal cortex organization for perceptual decision making. *PNAS*, May. <https://doi.org/10.1073/pnas.1522551113>
- Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, 20(4), 260–281. <https://doi.org/10.1016/j.tics.2016.01.007>
- Resulaj, A., Kiani, R., Wolpert, D. M., & Shadlen, M. N. (2009). Changes of mind in

- decision-making. *Nature*, 461(7261), 263–266.  
<https://doi.org/10.1038/nature08275>
- Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, 10(1), 1–11.  
<https://doi.org/10.1038/s41467-019-09075-3>
- Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*, 1(May).  
<https://doi.org/10.1017/pen.2018.16>
- Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry*, 1–9.  
<https://doi.org/10.1016/j.biopsych.2017.12.017>
- Sarafyazd, M., & Jazayeri, M. (2019). Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441). <https://doi.org/10.1126/science.aav8911>
- Schulz, L., Fleming, S. M., Dayan, P., Schulz, L., & Fleming, S. M. (2021). Metacognitive Computations for Information Search : Confidence in Control. *BioRxiv*, 1–35. <https://doi.org/10.1101/2021.03.01.433342>
- Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How Local and Global Metacognition Shape Mental Health. *Biological Psychiatry*, 18, 1–11.  
<https://doi.org/10.1016/j.biopsych.2021.05.013>
- Sepulveda, P., Usher, M., Davies, N., Benson, A., Ortoleva, P., & Martino, B. De. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *BioRxiv*, 2020.04.14.031971.  
<https://doi.org/10.1101/2020.04.14.031971>
- Shadlen, M. N., & Shohamy, D. (2016). Decision Making and Sequential Sampling from Memory. *Neuron*, 90(5), 927–939.  
<https://doi.org/10.1016/j.neuron.2016.04.036>
- Shea, N., & Frith, C. D. (2019). The Global Workspace Needs Metacognition. *Trends in Cognitive Sciences*, 23(7), 560–571. <https://doi.org/10.1016/j.tics.2019.04.007>
- Shekhar, M., & Rahnev, D. (2020). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, 1–12. <https://doi.org/10.1016/j.tics.2020.10.007>



- Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience*, 6(MAY), 1–20. <https://doi.org/10.3389/fnins.2012.00070>
- Travers, E., Fairhurst, M. T., & Deroy, O. (2020). Racial bias in face perception is sensitive to instructions but not introspection. *Consciousness and Cognition*, 83. <https://doi.org/10.1016/j.concog.2020.102952>
- Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron*, 96(2), 348-354.e4. <https://doi.org/10.1016/j.neuron.2017.09.006>



# Bridge:

## From inference to confidence.

This thesis argues for the role of subjective value in providing a comprehensive picture of procedural metacognition as a thermostat for decision coherence. By developing the models of both the function and the computation of metacognitive monitoring signals, we suggest that subjective value is essential to close the loop and provide a comprehensive understanding of metacognition. The second half of the thesis concerns empirical work that aims at providing new models for the computation of confidence reports in regard to subjective value.

The first half of the thesis was conceptual and took a multi-disciplinary approach in the aim to provide a complete picture of procedural metacognition: accounting for both its function as a behavioural and cognitive thermostat but also the computation of its monitoring signals. The first chapter aimed at providing the edges of the field: on one hand we illustrated what an autonomous agent might achieve by regulating the coherence of his behaviour, on the other hand, we reviewed the state of the research that strives at explaining how metacognition might perform such regulation of human cognition and behaviour. We highlighted between both ends the importance of accounting for subjective value to explain the workings of procedural metacognition. The second chapter reviews the literature in order to dive deeper on this gap and provides a landscape for metacognitive monitoring signals: focussing on executive functions that regulate behaviour as output, we define an array of metacognitive monitoring signals that might guide behavioural control. We then review the literature that defines how such monitoring signals are computed and most importantly the role of subjective value as input and explicit reports as monitoring signals to unlock these most advanced executive functions. Chapter 3 lastly looked at procedural metacognition from a different angle: how, on the cognitive side, does metacognition act as a thermostat to, in turn tune behaviour to

the context. While focussing on the literature in explicit confidence report, we highlight the ubiquitous role of metacognition as a tuning function for inference and how subjective value serves as a central input to inform the agent about the contextual reliability of her decision.

These last two chapters therefore highlight the central role of subjective value as input to regulate both cognition and behaviour through more or less explicit monitoring signals. Now focussing on the computation side, we propose to investigate empirically the more or less explicit role of subjective value as input for confidence reports. Mainly, Chapter 3 suggests that confidence signals are ubiquitously informed by subjective value which is formed by learning about the environment, one can also ask: in context with limited knowledge (*i.e.* before learning), what input informs confidence signals? In other words, if metacognition tunes cognition and behaviour thanks to its ubiquitous computation of subjective value that is built through learning about the context: in tasks where no learning has taken place and explicit knowledge is scarce, what kind of subjective value then serves as input to inform confidence reports about the adequacy of the decision?

Chapter 4 takes an empirical approach to investigate the role of heuristic cues, often implicit and shared amongst individuals, in informing agents about the possible reliability of their choices. Mainly, in such tasks with limited knowledge confidence reports were demonstrated to go hand in hand with popular opinions rather than accuracy itself. We test the nature of the reflective process by asking whether cues that guide the decision process (decision rule) are not also central to computing of confidence reports, therefore detaching metacognition from an accuracy tracker and describing instead a self-regulating thermostat. More specifically, we test whether reflective processes, both upon one's (metacognition) and others' (theory of mind) decisions, do not monitor the coherence with the rule guiding the decision, let it be explicit in tasks with clear knowledge or more implicit in this context. By doing so, we also test a common architecture to both theory of mind and metacognition as tracking coherence with oneself and with the group as well, both reflective processes eventually feeding into each other.

# Chapter 4:

## Confidence in art: the consensual illusion of accuracy.

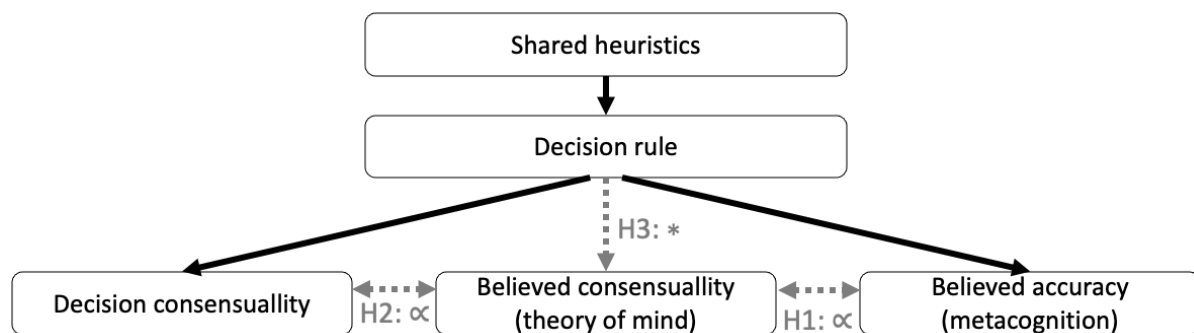
**Oriane Armand<sup>1,2,3</sup>, Joaquín Navajas<sup>4</sup>, Bahador Bahrami<sup>2,5</sup>, Ophelia Deroy<sup>2,3,6</sup>.**

1. Graduate School of Systemic Neurosciences, Ludwig Maximilian Universität. 2. Munich Center for Neurosciences 3. Institute for Philosophy of Mind, Ludwig Maximilian Universität, 4. Escuela de Negocios, Universidad Torcuato di Tella, 5. Institute of Cognitive Neuroscience, University College London, 6. Institute of Philosophy, University of London.

### **Abstract**

In tasks with limited knowledge about the correct decision rule (e.g. how to choose the best wine?), confidence was demonstrated to rely on social information when present (e.g. friend's advice) or to reflect decision consensuality otherwise (i.e. most people chose the same wine). While this link between confidence and consensuality was argued to rely on a shared heuristic decision rule (e.g. common belief that best wines have vintage labels), here we ask whether this link could not be explained by participants' inference of social information to cue their confidence. In other words, participants would be more confident in consensual decision not only as a consequence of a common decision rule, but because they would use their belief about others' behaviour (c.f. Abs. Fig. H1) and be correct in this inference (theory of mind ability, Abs. Fig. H2). Our experiment took place at the Tate Modern Gallery where lay participants (limited knowledge in N=49) guessed the price category to which 12 paintings in sale belonged to, and reported both their confidence and their belief about the consensuality of their answers. We replicated the finding that in this

task with limited knowledge, confidence was predicted by decision consensuality (rather than accuracy), suggesting the use of a shared heuristic for both cognitive and metacognitive processes. Our results revealed that believed consensuality was the best predictor of confidence amongst other predictors (H1), and furthermore, that the link between confidence and consensuality previously described was explained by the ability of participant to track accurately their decisions' consensuality (H2). Our results therefore extended the model of self-consistency by demonstrating that participants use the same heuristic rule not only to decide and reflect upon their own decisions but also to reflect upon others' behaviour. While we cannot conclude on causality within our experiment, our results suggest that metacognition and theory of mind might therefore share mechanism if not support each other in complex task by combining how we reflect upon our and other's behaviour.



**Abstract figure:** Could the link between decision consensuality and confidence be explained by an inference of the former to cue the latter? In tasks where participants have limited knowledge about the decision rule to follow, participants tend to use social information to cue their confidence, here we hypothesise that they (H1) infer about their decision's consensuality which could predict or inform confidence, (H2) that they accurately infer consensuality (theory of mind ability) and (H3) that the cues which define together decision consensuality and confidence also define believed consensuality.

## 1. Introduction

“Wait, I am not sure I chose the best wine.” While wine experts might have learned the accurate decision rule to follow when making and evaluating their decisions such as using a refined combination of vintage, domain and grapes, most of us only rely on an approximate heuristic rule such as the attractiveness of the bottle’s label. In such tasks with limited knowledge, it was demonstrated that, if present, participants often follow social information as to help them guide both their decisions (i.e. herding) and how confident they are about them. But in absence of social information, could participants use their theory of mind to inform their confidence about what others might do? Could the way in which we reflect upon the accuracy of our own decisions be inherently linked to the way we reflect about the decisions of others?

Confidence levels predict how agents invest in or correct their choices by capturing how reliable these are in regards to the decision rule. In the laboratory, confidence levels are studied for how they relate to the decision rule in different tasks. First, in perceptual tasks, choices are made with an explicitly instructed decision rule and confidence levels are observed to be best represented by the amount of evidence supporting the decisions. For instance participants would be asked to choose amongst two circles the one containing the most dots. In such task, confidence reflects the probability of the participant to have chosen correctly. It can be observed that the amount of evidence supporting the choice (e.g. the difference in dot number between both options) predicts both the probability of the choice to be correct and the level of confidence in this decision. This ubiquitous link suggests (amongst other parameters such as noise or cognitive fluency) that participants rely on the same cues (defined by the decision rule) to make their decision and monitor its reliability (Bang & Fleming, 2018; Pouget et al., 2016; Rouault et al., 2018). Secondly, this same link between the decision rule and confidence was also observed in value-based tasks. In these tasks, participants are not instructed about the cue on which to base their decisions but they define it subjectively given their own goal such as deciding to choose the snack which has either less calories or else has more continence in cacao. Again, when defined subjectively this time, it was observed that the decision

rule and its associated cue define confidence levels: the greater the difference in subjective value between the chosen and unchosen items, the greater the level of confidence (Boldt et al., 2019; Hertz et al., 2018). Therefore, not matter whether the decision rule is explicitly instructed or subjectively defined by personal experience, it is central to defining who the agent evaluate the reliability of her decision. But in tasks where participants have limited knowledge about the decision rule to follow, what defines confidence levels?

Self-Consistency Theory (SCT) of metacognition suggests that ultimately, confidence relies on the cues that, as defined by the decision rule, predict behavioural consistency (Koriat & Adiv, 2015). In simple terms, it means that what guides decisions (i.e. the decision rule) is also central to what guides decision monitoring (i.e. confidence levels). Empirically, the theory therefore suggests that, independently from stimuli uncertainty as studied in perceptual tasks, the uncertainty in the decision rule itself should predict both the average behavioural consistency and average confidence, and also predict how well the latter tracks the former. In special cases where the decision-rule is crystal clear such as when the rule is explicitly instructed (e.g. perceptual tasks) or that the agent is an expert at the task, behavioural consistency and confidence are mainly defined by the uncertainty in the stimuli (e.g. signal strength, noise, volatility...). The effect of the clarity of the decision rule on the links between confidence and consistency can be observed at two levels: within the individual with choices that are consistent over time are tagged with higher confidence, and within the group choices that are made consensually are tagged with higher confidence. In cases however where knowledge about the norm of accuracy is limited, it was observed that participants could consistently report high confidence in consensual decisions while these would be inaccurate (Koriat, 1976; Koriat et al., 2015). This link between confidence and consensuality rather than with accuracy was argued to be a marker of limited knowledge about the accurate decision rule and of the use of a shared heuristic within the cohort. But while we previously defined how in both perceptual and value-based tasks confidence relies on cues defined by the decision rule, it remains unknown what confidence actually relies on when knowledge it thus limited. Indeed, this definition



of confidence portrays it as an ubiquitous measure of the choice consistency both within the individual and within the group. But beyond this descriptive observation of a relation between confidence monitoring and group consensuality, one can ask whether when using a heuristic decision rule the social human brain would rely on similar mechanisms and cues to monitor its own decisions and the group's decisions. In other words, in context where one's grip on accuracy is limited, can different reflective processes such as metacognition and theory of mind function independently or do they share the same limited evidence? How does our monitoring of our own and of other's decisions influence each other and what does it mean about both functions?

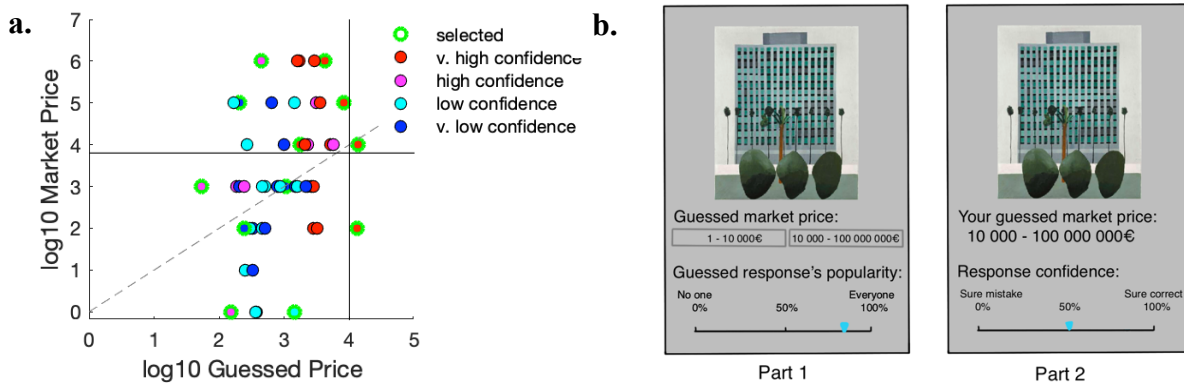
Both evolutionary and developmental theories define some co-dependence in the emergence of metacognition and theory of mind, respectively across species and within individuals (Heyes et al., 2020). In both cases, an agent with any of these abilities is able to have multiple hypothetical models of the world in mind: either one's actual behaviour against one's expected behaviour (*i.e.* according to decision rule); or one's decision rule against other's decision rule (Carruthers, 2009). Furthermore, beyond the similarity in both these monitoring functions as being comparative, previous literature has highlighted the intimate relation between social information and confidence judgments. Specifically, in contexts where knowledge is limited, it was demonstrated that lack of evidence on which to evaluate one's decision tends to be compensated by the use of social evidence (De Martino et al., 2017; Pescetelli & Rees, 2016). Therefore, beyond computational similarities which could result into output similarities, it appears that one type of information could leak into the other, resulting into both processes not only working in parallel but also working together for a common monitoring function. Here we ask whether in contexts where participants have limited knowledge, an individual who is confident in her choice is also likely to believe that this choice is consensual. In other words: would the cues on which one rely to make and evaluate a decision be common to the cues on which one would rely to infer about other's behaviour? We expect that both reflective inference and evaluation of one's choice would go hand in hand. In other words we ask whether, beyond the observation that confidence and

consensuality go hand in hand, both inference processes could go hand in hand and eventually feed into each other from a subjective perspective. More specifically, we hypothesise that when participants reflect about accuracy of their decisions, they would use similar cues and therefore correlate their evaluation about their decisions' consensuality. We discuss how this metaknowledge, when shared, can therefore define similar confidence profiles across individuals and enable them to infer about their decisions in relation to the group's behaviour.

## 2. Methods

### *Stimuli*

To study the links between how participants reflect upon their decision accuracy and consensuality, we used a domain with limited and shared knowledge where decision confidence and consensuality go hand in hand: the art market (Prelec et al., 2017). In our task, 51 paintings from the auctions site *Ketterer Kunst* (<https://kettererkunst.com>) and *Christie's* (<https://www.christies.com>) were pre-selected for having a market price as estimated by experts within a year before the experiment ranging from 1 to 100 000 000 euros. The price range was considered for spanning across several log<sub>10</sub> units to create even price categories taken into consideration that participants appear to represent large number on a log rather than linear scale (observed in our previous pilots and as mentioned in literature (Dehaene, 2003)). The paintings also all had similar dimensions (about 50 by 60 cm) and belonged to various time periods and artistic movements (realism, impressionism, abstract art...). For an online pilot (with testable.org and Amazon Turk), they were presented to 32 participants who had to guess the market price given by experts and report their levels of confidence in having estimated correctly. From these pre-selected paintings, 12 paintings (*c.f.* supplementary material) were selected to obtain a list with market prices being on average correctly or incorrectly guessed and resulting on average in low or high confidence (Fig 1a). The resulting bank of stimuli therefore presented participants with a list of items on which their knowledge was limited by presenting choices of various difficulty for our study on confidence levels.



**Figure 1: Stimuli selection and experimental procedure.** **a** Pilot preselection of 12 paintings amongst 51 presented to participants (N=32) who had to guess their market price and report their confidence in the accuracy of their estimates. For each painting the log<sub>10</sub> average gussed price is plotted against the log<sub>10</sub> category of its actual market price as presented for auctions. The black lines define the boundary between both price options namely at 10 000 euros in the final task. The dotted line is the identity, whereby items on the line are on average correctly estimated, those above and below are respectively over and under estimated. The colour coding represents the quartiles of average z-scored confidence level across participants, and the green items were selected for the final task for representing these various dimensions of interest. **b.** The experiment comported two parts which were interchangeably presented across participants for either asking believed consensuality or confidence first. In first part participants saw the 12 paintings in a random order and had up to 15 seconds to both chose a price category and report belief about the decision. In the second part participants saw all the paintings again in random order and where asked to make the second judgment about their decision which was reminded to them.

### *Procedure*

As part of the Tate exchange program, participants were visitors of the Tate Modern Gallery in London who voluntarily took part in a short experiment. Participants were asked whether they would like to take part in a 10 minutes experiment about the art market where they would have to answer questions about paintings in sales for

auctions. If so, they were given a google tablet with the link to the experiment designed in Qualtrics. They were verbally explained that the experiment would be in 2 parts: In the first part, they would have to guess price range of paintings in sale for auction as defined by expert and – either rate their confidence in having guessed correctly from 0 to 100% certain; or how many people they thought decided as they did from 0 to 100%. In the second part of the experiment, participants would have to answer the complementary reflective question which they were not attributed in the first part. When considering the consensuality of their decisions, participants were asked to infer the decisions of other participants, random visitors of the museum with no specific expertise in the art market. Finally, participants were informed that to unsure their focussed attention, the question on the screen would disappear within 15 seconds if not answered.

Participants then had to read and agree on a online consent form, respond to a few questions about themselves and read general instructions in great details about the task at hand. In the first part of the experiment, participants were presented with each painting successively and asked to guess whether its price (as evaluated by experts) was within the 1 to 10k euros or the 10k to 100m euros category. Following this decision participants were asked to reflect about it: half of the participants were directly asked to estimate their confidence in their decisions and then their belief about the popularity of their decision in the second part, and other half of participants were presented with the reversed order (Fig 1.b). Both of these scales were presented in % from 0 to 100 with an indent every 10%. The confidence scale had for labels “0% sure error” and “100% sure correct”; the popularity scale had for labels “0% no one” and “100% everyone”. In the second part, each painting and decision was shown again to the participant for them to perform the second inference about their decision. In both parts, paintings were presented in random order and participants had a maximum of 15 seconds to answer question as a gentle incentive to focus on the task.

### ***Participants***

A total of 51 visitors volunteered to take part in this short experiment. Amongst them, many young participants in their teens were keen on participating in the study

(skewed age distribution: age mean 19.02, std= 14.73). Participants were also asked about their level of expertise in the art market based on 3 questions: do they buy or sell art, are they in the art business and did they study art (average score: 0.53/3 SE=0.79). Two participants were excluded for not having completed the entire experiment.

### ***Data processing***

Trials with at least one of three response lacking (decision, confidence, believed consensuality) were removed (53 trials out of 588). For each painting, the objective consensuality was calculated across the N=49 participants as a ratio between both answer (scale in percentage as for belief popularity and confidence) and respectively attributed to each participant's response. Responses, response times, confidence levels, beliefs about responses' consensuality and objective consensuality of the responses were z-scored within participants. Finally all parameters were z-scored across participants to ensure their fair comparison in multiple regressions.

## **3. Results**

We asked whether in tasks with limited knowledge, participants' reflection about the accuracy and consensuality of their choices would go hand in hand. For 12 paintings, volunteers chose amongst two price categories the one they believed to contain the market price and reported both their confidence in the accuracy of their choices and their belief about their consensuality.

### ***Confidence and consensuality***

First of all, we controlled for the assumption that, as in tasks with limited knowledge, our participants decisions and confidence levels were linked by the norm of consensuality rather than accuracy. We observed that despite variance, some participants performed well above chance level when choosing the paintings' price categories (Sup fig 1: mean= .58, std=.14). For each painting, response consensuality was calculated from our cohort's choices amongst both price categories. In this

study, we used mixed models to capture the predictors of confidence across the small number of trials per participants while keeping subjective noise into account as a random effect. Replicating previous findings in the art market, our lay participants' confidence levels were better predicted by their response consensuality than accuracy (sup fig 3.a: consensuality:  $\beta = .23$  ,  $SE = .05$   $z = 4.75$ ,  $p < .001$ ; accuracy:  $\beta = -.04$  ,  $SE = .05$   $z = -0.82$ ,  $p = 0.41$ ; sup fig 3e:  $BIC(\text{consensuality}) = 1518$  ;  $BIC(\text{accuracy}) = 1541$ ). Altogether, these results suggest that participants have a limited knowledge about the art market as their decisions and confidence levels are not ruled by the norm of accuracy (as unique predictor in model 1 sup fig3e:  $\beta = .06$ ,  $SE = .04$ ,  $z = 1.41$ ,  $p = 0.16$ ). But that their choices are still guided by a consistent decision rule as a heuristic shared within the cohort. In the following part, we ask whether participants are subjectively aware of this observed link between theirs and others decisions when monitoring their accuracy.

### ***Illusion of accuracy***

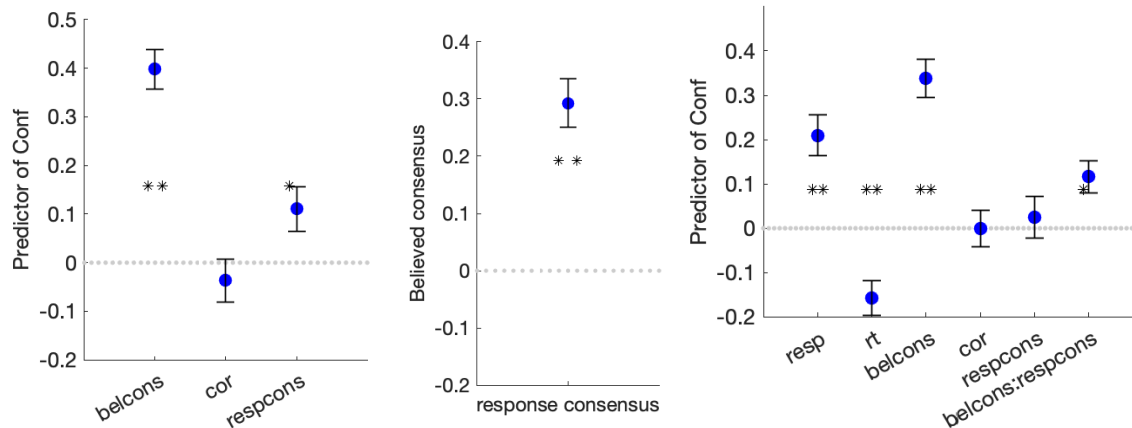
Beyond the observed link between confidence and consensuality as relying on a shared decision rules, here we ask whether participants subjectively track the group behaviour to cue the monitoring of their own decisions' accuracy. To link this objective consensuality with the subjective confidence monitoring, we therefore asked participants to report their belief about their decisions consensuality to see whether it was intertwined with their beliefs about their decision's accuracy. We compared our three main candidate predictors of confidence in a mixed model and found that belief consensuality was an even better predictor of confidence than the previously discussed observed consensuality (fig 2a: belief consensuality:  $\beta = .40$ ,  $SE = .04$   $z = 9.72$ ,  $p < .001$ ; response consensuality:  $\beta = .11$ ,  $SE = .05$   $z = 2.37$ ,  $p = 0.02$ ). Furthermore, amongst all tested models of confidence, the model with the unique predictor of belief consensuality was the best (sup fig3e:  $BIC = 1437$ ). This link between response consensuality and both reflective processes (about both consensuality and accuracy) suggests that the cohort decision and reflective processes must share cues defined by the heuristic decision rule (abstract's figure). Further than this common denominator to decision and reflective processes, we

then asked whether participants could reliably track their response consensuality which we here call a Theory of Mind (ToM) ability.

### *Inferring consensuality*

We previously observed that participants do not demonstrate metacognitive ability in the art market by not managing to discriminate their correct from incorrect decisions with their levels of confidence. Instead these latter were predicted by decision consensuality and the participants beliefs about their consensuality. To finish bridging confidence with consensuality by belief consensuality, we now asked whether participants demonstrated a ToM ability by discriminating accurately consensual from non-consensual decisions with their beliefs. Indeed, we observed a significant ability of participants at inferring whether their decisions aligned with other participants' decisions (Fig 2b:  $\beta=.23$ ,  $SE=.05$   $z=4.75$ ,  $p<.001$ ). Therefor in this task with limited knowledge about accuracy, participants appear to lack metacognitive ability but to have a ToM ability which accurately monitors the group's behaviour in this two alternatives choices.

Finally, we asked whether the cues on which ToM ability relied to correctly infer consensuality also predicted confidence levels. While a model with this interaction was not a better predictor than the more parsimonious model with belief consensuality in its own (sup fig 3e: model 7 BIC (respcons+belcons)= 1456; model 8 BIC (respcons\*belcons)= 1472) we observed that when accounting for cues supporting ToM ability, the link between response consensuality and confidence disappeared (fig 2c: belcons:respcons:  $\beta=.12$ ,  $SE=.04$   $z=3.15$ ,  $p<.05$ ; respcons:  $\beta=.02$ ,  $SE=.05$   $z=0.52$ ,  $p=.60$ ; see sup fig 3c,d for effect of ToM on confidence in more parsimonious models). In other words, the link between response consensuality and confidence (fig 2a) can be significantly explained by the link between response consensuality and believed consensuality (fig 2c). While our experiment does not permit us to conclude on a causal effect of belief consensuality as leaking into confidence more than the other way around, it can nonetheless be confirmed that the cues on which the cohort relies to decide is common with both ToM and metacognitive inferences.



**Figure 2: reflexive functions upon accuracy and consensuality and their predictors a.** mixed model (model 6 in Sup Fig 2e) of confidence levels with repones correctness, response consensuality and believed consensuality **b.** mixed model of believed consensuality as predicted by objective consensuality, also referred to in this article as theory of mind ability **c.** mixed model (model 12 in Sup Fig 2e) of the predictors of confidence with respectively response category (or magnitude), response time, believed consensuality, response correctness, response consensuality and the interaction between believed consensuality and objective consensuality. Significance of predictors is indicated by \* for  $p < .05$  and \*\* for  $p < .001$ .

### ***Metacognitive cues***

Lastly we computed a model with our predictors of interest together with other known predictors of confidence to account for all their effects in a unique model. Interestingly, in this full model, our belief consensuality parameter was still the best predictor of confidence levels amongst other candidates (Fig2c:  $\beta = .34$ ,  $SE = .04$   $z = 7.81$ ,  $p < .001$ ). As previously defined, this full model still presents the cues of response consensuality which contribute to confidence as captured by the participants ToM ability. The two other parameters added to the model were response time suggesting that slow response are generally rated with lower confidence ( $\beta = -.16$ ,  $SE = .04$   $z = -3.99$ ,  $p < .001$ ). Lastly, in this inherently evaluative task, we found that response magnitude (or price category) was predictive of confidence levels ( $\beta = .21$ ,  $SE = .05$   $z = 4.49$ ,  $p < .001$ ). This result replicate previously observed results



whereby detection of cues for high evaluation also support judgment of confidence (Fleming, 2020; Lebreton et al., 2009).

### ***Beyond 2AFC***

Lastly, we asked whether this effect of consensuality on confidence was limited to tasks where the population is clearly divided in two -a majority and minority- by the presence of only two options. We designed three additional tasks, respectively of 3AFC, 4AFC and 6AFC, in which the same paintings' prices were distributed evenly across price categories (Sup Table 1). We conducted post hoc analyses to evaluate the effect of the increasing number of alternatives on the predictors of confidence as in our full model above (Sup Fig 4a). While not the most parsimonious (Sup Fig 3g: model 22 BIC=4868; best model 4 (response magnitude) BIC=453), this model demonstrated first of all that across all tasks, consensuality remained a better predictor of confidence than accuracy (respectively Sup Fig 4a: consensuality:  $\beta = .07$ ,  $SE = .03$ ,  $z = 2.42$ ,  $p < .05$  and accuracy:  $\beta = -.02$ ,  $SE = .02$ ,  $z = -.93$ ,  $p = .35$ ). These replications suggest that, once again, participants use a shared heuristic to guide their decisions and metacognition. Furthermore, as in 2AFC, our first hypothesis was confirmed across all tasks by demonstrating that belief consensuality was still a better predictor than observed consensuality ( $\beta = .25$ ,  $SE = .03$ ,  $z = 9.27$ ,  $p < .001$ ). Therefore, we can still observe the link between both reflective functions (believed consensuality and believed accuracy) in more than the 2AFC. Regarding this first hypothesis linking inference about accuracy to inference about consensuality, we tested the (linear) effect of increasing number of alternative choices on this relation. Across all tasks we did not observe that increasing number of options linearly broke down this link between confidence and believed consensuality ( $\beta = -.04$ ,  $SE = .03$ ,  $z = -1.75$ ,  $p = .08$ ). We then tested the difference between 2AFC and respectively 3 and 4 AFC. Our central hypothesis here was that the loss of the polarity present in the 2AFC would make consensuality an irrelevant proxy for accuracy as the task structure offers a neutral middle ground where some participants do not take a side more than the other. In that reasoning the 4AFC is not a structure presenting even more neutral ground (linear increase) but instead recovering a polarity with a "rather low" or "rather high" price categories. Our independent analysis of an effect of number of

alternative between 2 and respectively 3 and 4 AFC did indeed reveal that the link between participants' believed accuracy and believed consensuality significantly broke down (Sup Fig 3b:  $\text{belcons}*\text{task}$ :  $\beta=-.11$ ,  $\text{SE}=.05$ ,  $z=-2.02$ ,  $p<.05$ ; Sup Fig 3c:  $\text{belcons}*\text{task}$ :  $\beta=-.11$ ,  $\text{SE}=.05$ ,  $z=-2.24$ ,  $p<.05$ ).

As previously, we then looked at the other side of the equation by looking at the link between believed consensuality and observed consensuality: ToM ability. Our reasoning was that in either increasing number alternative or at least when polarity is lost with the task structure (mainly 3AFC) participants would lose their ability to accurately infer the cohort's behaviour. We observed that ToM ability decreased from 2AFC to 3AFC (Sup Fig 3b:  $\beta=.29$ ,  $\text{SE}=.04$ ,  $z=6.98$ ,  $p<.001$ ;  $\beta=.11$ ,  $\text{SE}=.05$ ,  $z=2.20$ ,  $p<.05$ ) but was improved when polarity of alternative reappeared in the 4AFC structure ( $\beta=.22$ ,  $\text{SE}=.05$ ,  $z=4.83$ ,  $p<.001$ ). The small number of volunteer we managed to recruit for the 6AFC does not permit us to reliably conclude on the obtained results for this task.

Lastly, we looked at whether the cues shared by confidence and consensuality were indeed captured by ToM. We focused on a more parsimonious model with only significant factors from the full model (Sup Fig 3c). We observe that indeed evidence for ToM contributes to confidence levels ( $\beta=.09$ ,  $\text{SE}=.02$ ,  $z=4.39$ ,  $p<.001$ ), however we can notice that accounting for this evidence in 2AFC explained the link between confidence and observed consensuality, this latter remains significant when the number of option increase here ( $\beta=.06$ ,  $\text{SE}=.03$ ,  $z=2.32$ ,  $p<.05$ ). In other words, while in 2AFC cues linking confidence to consensuality were also picked up by belief consensuality, this relation is not as strong in tasks with more alternatives. We note however that in our model where the effect of task is encoded linearly by its number of alternatives, we do not observe a significant effect of alternative number on these links between confidence and believed consensuality.

Taking into account that response magnitude appears to contribute about as much to confidence that belief consensuality in these more complex tasks (Sup Fig 3c respectively  $\beta=.24$ ,  $\text{SE}=.03$ ,  $z=8.02$ ,  $p<.001$  and  $\beta=.27$ ,  $\text{SE}=.03$ ,  $z=9.79$ ,  $p<.001$ ), we then looked at the interaction of this parameter within our model aiming to explain the link between confidence and consensuality through belief consensuality (Sup

Fig 3d). In other words, we asked the question whether the cues supporting both the cohort consensual behaviour and confidence could (if not shared with ToM which could be challenged by complex task structures) be captured by the magnitude of the response itself (as suggested by detection theories previously mentioned). The rational being that while inferring other's behaviour might become irrelevant or too complex to compute when majority-minority dissolves in numerous alternatives, the decision rule defining both cohort consensuality and subjective confidence could be better captured by the magnitude of the response itself. We observed in this next model that confidence' link with consensuality and believed consensuality was better explained by their shared features with response magnitude (Sup Fig 3d: resp\*respcons:  $\beta = .07$ ,  $SE = .02$ ,  $z = 2.99$ ,  $p < .05$ ; resp\*belcons  $\beta = .08$ ,  $SE = .02$ ,  $z = 3.30$ ,  $p < .001$ ; while respcons\*belcons drops:  $\beta = .03$ ,  $SE = .02$ ,  $z = 1.20$ ,  $p = .23$ ). In other words, in more than 2AFC structures, ToM does not share as much with metacognition but cues that predict high magnitude responses also predict the response to be consensual, believed to be consensual and expressed with high confidence.

#### 4. Discussion

In this study, we asked whether the link observed in limited knowledge tasks between confidence and choice consensuality was also subjectively aware by having a believed consistency relating to both these factors. We observed indeed that lay participants tended to link their beliefs about their decisions' consensuality and accuracy, while also going hand in hand with the decision's consensuality amongst our cohort. Our analysis reveal that these three parameters seem to rely on the same cues to guide both the decision consensuality across the cohort and both these reflecting processes. In a task with limited knowledge, these similarities between monitoring the accuracy of one's decision (metacognition) and its consensuality (theory of mind) could be explained either by a leak between both computations or by foundational similarities in their respective computations.

While in tasks with limited knowledge social information can inform our confidence in our choices' accuracy (De Martino et al., 2017), here we asked whether in lack of

such evidence, participants would link their beliefs about their choices' consensuality to their own confidence levels. Similarly, in our task, the decision rule was defined by neither of explicit instruction or subjective knowledge, instead, we replicated conditions where behaviour and metacognition appear to rely on a heuristic decision rule (Koriat et al., 2015; Prelec et al., 2017). Our results indeed demonstrated a link between lay participants beliefs about their response consensuality and accuracy. It remains unclear whether however, such as social information feeding into confidence, this belief about consensuality informs causally confidence in a causal manner.

Besides the prediction of confidence by believed consensuality, our study also accounted for other markers of confidence. Indeed, response time (as a marker of cognitive fluency) is a widely acknowledged heuristic for confidence levels (Shekhar & Rahnev, 2020). Besides, in our present study, we found a very strong effect of response magnitude on confidence whereby paintings guessed as expensive were consistently rated with higher confidence than lower price responses. This effect reflects a commonly known effect where in estimation tasks, participants are highly confident when evidence supports one option against another but that ambiguous items lead to lower confidence (Lebreton et al., 2015). Similarly we can interpret the present results as participants knowing better when they recognise a cue for expensive painting but both deciding for lower price and lower confidence in absence of such striking cue. These findings aligns with recent research in metacognition studying confidence of signal absence whereby participants, while being able to know when signal was absent, did use lower confidence levels to report absent that present signals (Fleming, 2020).

Beyond this similarity between evaluation of decisions' accuracy and consensuality, we asked whether both of these reflexive process could rely on the same evidence that links them to the cohort consistent behaviour. According to SCT, this would therefore imply that both reflective functions use the same heuristic decision rule to monitor the consistency of choices. Our results revealed that the link between consensuality and confidence relied on the same cues as the link between consensuality and believed consensuality. While, as previously discussed, this

relation between both reflective evaluations could be explained by a causal leak such as believed consensuality serving as heuristic to inform believed accuracy (*i.e.* confidence), an alternative explanation takes root in the rich evolutionary and developmental literature linking both these reflexive processes (Fleming, 2021). Indeed, both abilities provide the agent with the ability to compare her own thinking and behaviour to alternative ones thereby comparing two (or more) possible models of the world. In this regard, SCT is insightful as it suggests that metacognition, as a reflexive process, uses a norm of consistency to evaluate choices. It suggests that this norm is task specific and guides both the choices and their monitoring. Depending on the task at hand, this decision rule as norm could therefore be one's subjectively established preferences and goals or even explicitly instructed targets in perceptual tasks. In socially relevant tasks where the goal might be to fit in a social group by sharing its values, the decision rule might also be to follow consensuality. In that sense, metacognition or reflective processes in general can be seen as a same evaluative process within which the norm changes based on the task at hand. This idea of a second order cognition monitoring different types of reliabilities to guide the agent as with a unique compass can be seen as a parallel to theories of consciousness (Shea & Frith, 2019). How theory of mind and metacognition can come apart to track distinct norms could be studied in a learning task where the agents would progressively learn the norm of accuracy to be different from their initial heuristic. More simply, conducting the task with art experts could also be informative about the dissociations of the concepts of consensuality and accuracy when reflecting upon one's choice. This entanglement between metacognitive evaluation and the inference about group behaviour could also be studied in other culturally relevant matters such as politics and morality which are also central to cultural identity. It could be argued that the ambiguity of the concept of accuracy in the present study could make it an exception where theory of mind could serve as a general heuristic for confidence, even though the task was performed in isolation and focusing on self-monitoring.

We then challenged the relevance of monitoring consensuality in this task by breaking down the majority-minority split that is proper to 2AFC tasks. We expected

that by presenting the same task and paintings with increasing number of alternatives to different groups of participants, the link between believed popularity and confidence would break down. Indeed we hypothesised that if the task structure would not provide a simple divide of the population, participants would rely less on their theory of mind as a proxy to infer about the accuracy of their responses. We found that confidence could still be predicted by believed consensuality, but that increasing the number of alternative beyond 2 did affect this link between both reflective functions. Furthermore, theory of mind ability appeared to decrease in the presence of more than two alternatives. Instead, the link between confidence and consensuality in more than 2AFC seem to rely on cues that predict response magnitude rather than its believed consensuality. In tasks with limited knowledge and finer estimate to provide, the metacognitive process could therefore be seen as sharing parallels with detection theory than theory of mind. Indeed while the cohort might split in less trackable groups with increasing number of options, the strength of evidence guiding response magnitude could be a stronger predictor of confidence levels (Fleming, 2020; Lebreton et al., 2015). Although the present experiments provide a useful insight on the basis of reflection in ambiguous tasks, it can be noted that these short studies conducted on volunteers in an open public space would benefit from replications and more within participant trials to strengthen the present findings.

Confidence is argued to be a common currency of behavioural reliability that has for pragmatic role to guide subsequent behaviour. Several areas of research seem to support this multi facet computation and role. First, self-consistency theory of metacognition suggests that beyond tracking cues for decision accuracy, confidence has a wider role of monitoring cues for behavioural consistency in various types of tasks (Koriat & Adiv, 2015). Secondly, Global workspace theory also suggests that a monitoring of input reliabilities is a central pillar on which relies the unifying of various currencies into a singular conscious experience and guidance of behaviour (Shea & Frith, 2019). Third, the literature in reinforcement learning has also recently been developing more complex models of confidence computations that demonstrates its multi-facet role in refining learning and behavioural flexibility (Lak

et al., 2017; Sarafyazd & Jazayeri, 2019; Sepulveda et al., 2020). Lastly, the integration of social information into confidence levels also painted its computation as multi-facet common currency of decision reliability (De Martino et al., 2017). Here, we demonstrated that in tasks with limited knowledge confidence seems to rely on (or go hand in hand with) a new cue: inferred consensual. In the context of learning, it would be interesting to see how the computation of confidence evolves from the lay participant with limited knowledge to an expert with refined (or weighted) use of different sources of uncertainty. This calibration of confidence levels to different sources of uncertainty would make it a central anchor of behavioural flexibility, to match behavioural flexibility to environments with volatile or probabilistic reward contingencies (Boldt et al., 2017).

To conclude, we found that the link between confidence and consensuality, as predicted by self-consistency theory, revolved around the individual ability to infer accurately about their answer's consensuality and that this reflective process shared grounds with metacognitive monitoring. We discuss the implications of these findings for their meaning in the ability of participants to use a same task-specific decision rule both for decision-making and reflection, and the shared mechanisms between the processes of metacognition and theory of mind.

### **Acknowledgments**

We wish to thank the staff at the Institute of Philosophy of University of London and at the Tate Modern for having organised the event and all the volunteer participants who took part in the experiment. We also wish to thank Mattia Gallotti and the participants of the Human Mind conference that took part at Churchill College, Cambridge, for their feedback on the previous version of the experiment.

### **Data availability:**

Data from the 4 experiments can be downloaded online at the following link:

<https://github.com/orarmand/Confidence-in-art>

## References

- Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 6082–6087. <https://doi.org/10.1073/pnas.1800795115>
- Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, *2019*(1), 1–18. <https://doi.org/10.1093/nc/niz004>
- Boldt, A., de Gardelle, V., & Yeung, N. (2017). The impact of evidence reliability on sensitivity and bias in decision confidence. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(8), 1520–1531. <https://doi.org/10.1037/xhp0000404>
- Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*(2), 121–182. <https://doi.org/10.1017/S0140525X09000545>
- De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. *Journal of Neuroscience*, *37*(25), 6066–6074. <https://doi.org/10.1523/JNEUROSCI.3880-16.2017>
- Dehaene, S. (2003). The neural basis of the Weber-Fechner law: A logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4), 145–147. [https://doi.org/10.1016/S1364-6613\(03\)00055-X](https://doi.org/10.1016/S1364-6613(03)00055-X)
- Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of Consciousness*, *2020*(1), 1–9. <https://doi.org/10.1093/nc/niz020>
- Fleming, S. M. (2021). *Know Thyself: The Science of Self-Awareness*. Basic Books.
- Hertz, U., Bahrami, B., & Keramati, M. (2018). Stochastic satisficing account of confidence in uncertain value-based decisions. *PLoS ONE*, *13*(4), 1–23. <https://doi.org/10.1371/journal.pone.0195399>
- Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences*, *24*(5), 349–362. <https://doi.org/10.1016/j.tics.2020.02.007>



- Koriat, A. (1976). Another look at the relationship between phonetic symbolism and the feeling of knowing. *Memory & Cognition*, 4(3), 244–248.  
<https://doi.org/10.3758/BF03213170>
- Koriat, A., & Adiv, S. (2015). The self-consistency theory of subjective confidence. *Oxford Handbook of Metamemory*, 1(June), 1–25.  
<https://doi.org/10.1093/oxfordhb/9780199336746.013.18>
- Koriat, A., Adiv, S., & Schwarz, N. (2015). Views That Are Shared With Others Are Expressed With Greater Confidence and Greater Fluency Independent of Any Social Influence. *Personality and Social Psychology Review*, 20(2), 176–193.  
<https://doi.org/10.1177/1088868315585269>
- Lak, A., Nomoto, K., Keramati, M., Sakagami, M., & Kepecs, A. (2017). Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision. *Current Biology*, 27(6), 821–832.  
<https://doi.org/10.1016/j.cub.2017.02.026>
- Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, 18(8), 1159–1167. <https://doi.org/10.1038/nn.4064>
- Lebreton, M., Jorge, S., Michel, V., Thirion, B., & Pessiglione, M. (2009). An Automatic Valuation System in the Human Brain: Evidence from Functional Neuroimaging. *Neuron*, 64(3), 431–439.  
<https://doi.org/10.1016/j.neuron.2009.09.040>
- Pescetelli, N., & Rees, G. (2016). *The Perceptual and Social Components of Metacognition*. 145(8), 949–965.
- Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3), 366–374. <https://doi.org/10.1038/nn.4240>
- Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, 541(7638), 532–535.  
<https://doi.org/10.1038/nature21054>
- Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*, 1(May).

<https://doi.org/10.1017/pen.2018.16>

Sarafyazd, M., & Jazayeri, M. (2019). Hierarchical reasoning by neural circuits in the frontal cortex. *Science*, 364(6441). <https://doi.org/10.1126/science.aav8911>

Sepulveda, P., Usher, M., Davies, N., Benson, A., Ortoleva, P., & Martino, B. De. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *BioRxiv*, 2020.04.14.031971.

<https://doi.org/10.1101/2020.04.14.031971>

Shea, N., & Frith, C. D. (2019). The Global Workspace Needs Metacognition. *Trends in Cognitive Sciences*, 23(7), 560–571. <https://doi.org/10.1016/j.tics.2019.04.007>

Shekhar, M., & Rahnev, D. (2020). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, 1–12. <https://doi.org/10.1016/j.tics.2020.10.007>

## 5. Supplementary material

### 5.1 Supplementary methods for more than 2 alternative forced choice

#### *Procedure*

A total of four experiments were conducted at the Tate Modern Gallery, each experiment performed by different individual presented either 2, 3, 4 or 6 alternative. The procedure was similar in all tasks as defined in the above paper.

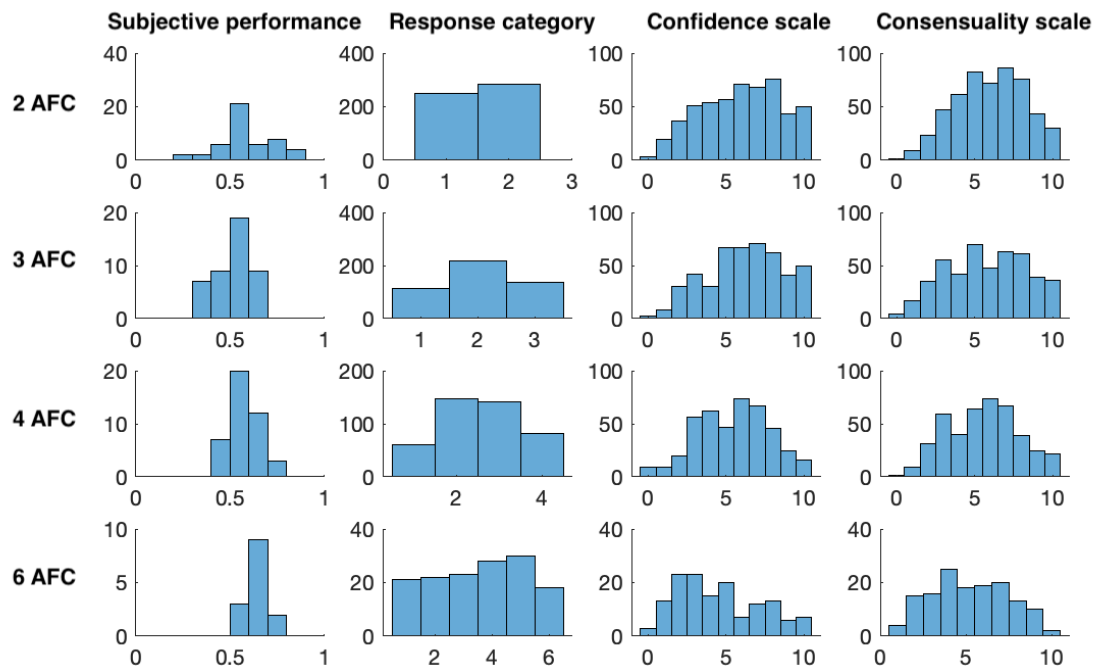
Price €	>1	>10	>100	>1 k	>10 k	>100 k	> 1m	>10 m
Log <sub>10</sub>	0	1	2	3	4	5	6	7
2 AFC	1			2				
3 AFC	1		2		3			
4 AFC	1	2		3		4		
6 AFC	1	2	3	4	5	6		

**Supplementary Table 1: Distribution of the painting selection for their market price to span equally across log<sub>10</sub> price categories of 2,3,4 and 6 AFC tasks.**

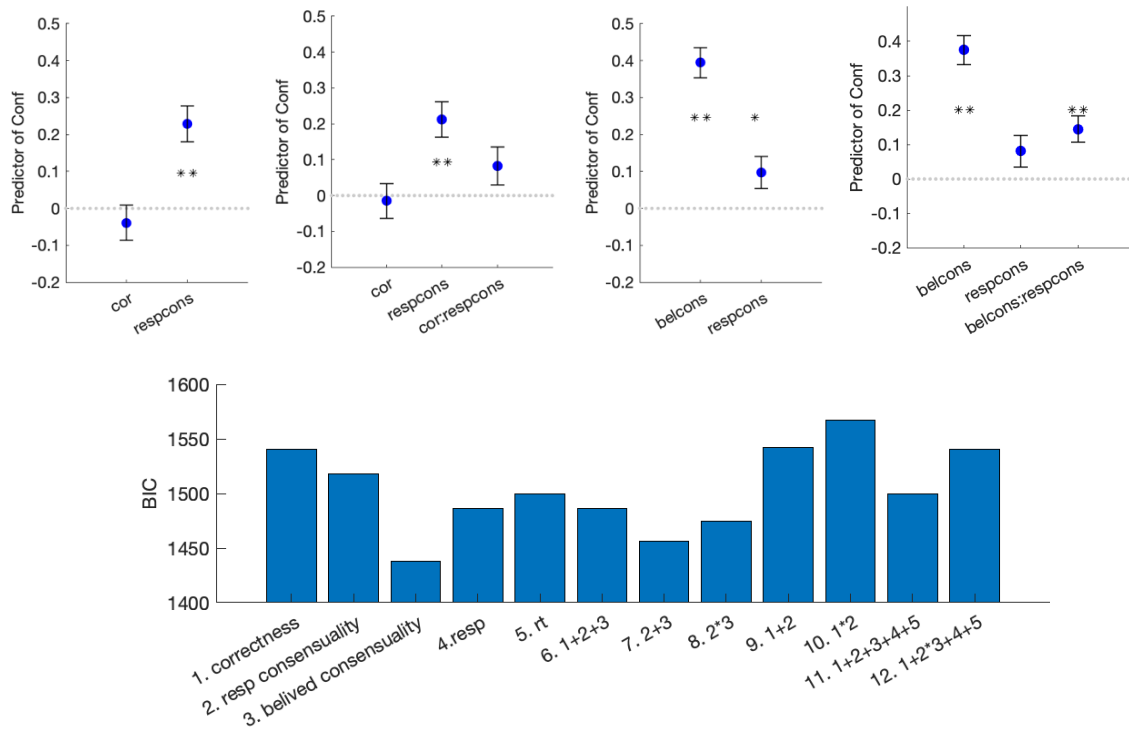
#### *Participants*

After the 2AFC task reached 50 participants, volunteers were selected to be older. In the 3AFC task 49 volunteers took part amongst which 2 were excluded for unfinished task and 3 for lack of variety in an answer type (which prevents the use of z score normalisation) (mean age=20.17, se=14.37, expertise average=0.7, se=0.95). In this task, 58 out of 528 trials were excluded for having a missing value. In the 4 AFC task, 49 participants took part, 3 were excluded for unfinished task, then 4 were excluded for lack of response variability (73 out of 504 trials were removed, participant age mean 22.2, SE=17.2, expertise score 0.53/3 SE=0.81). In the 6AFC task, 16 participants took part and 2 were excluded for lack of response variability (26 trials excluded out of 168, mean age=37.3, SE=14.5, expertise score=0.69/3, SE=0.87).

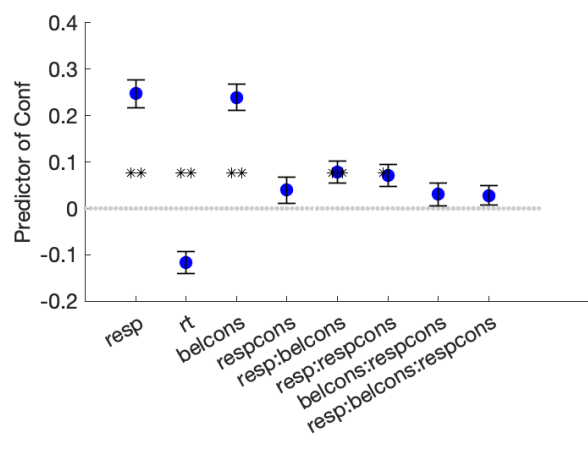
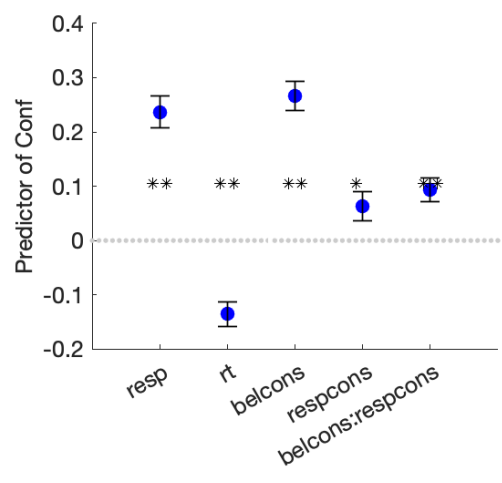
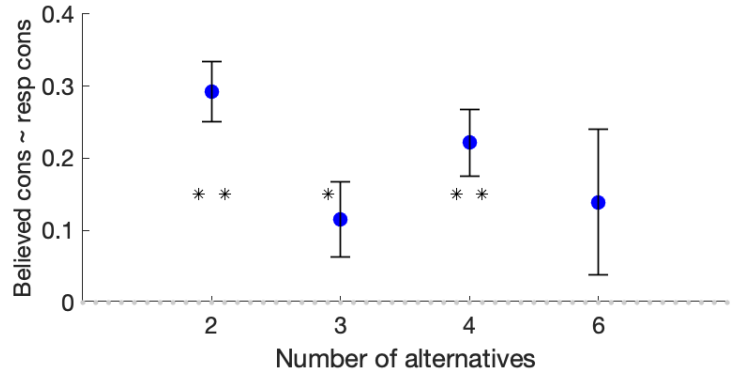
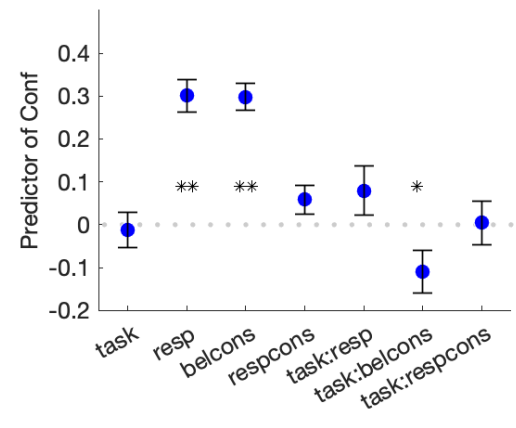
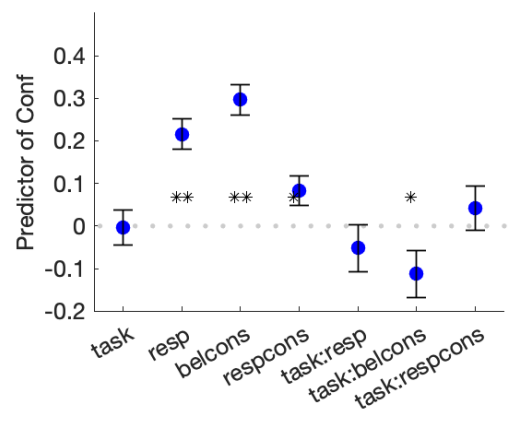
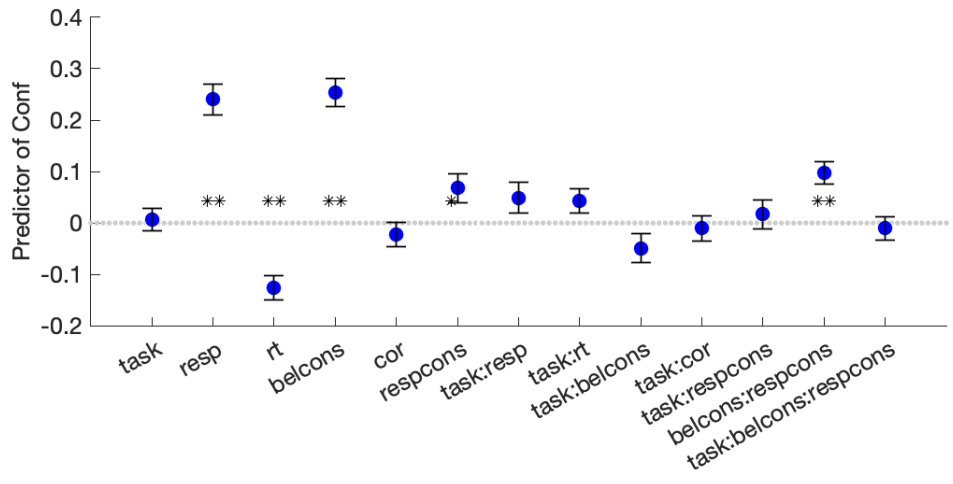
## 5.2 Supplementary figures

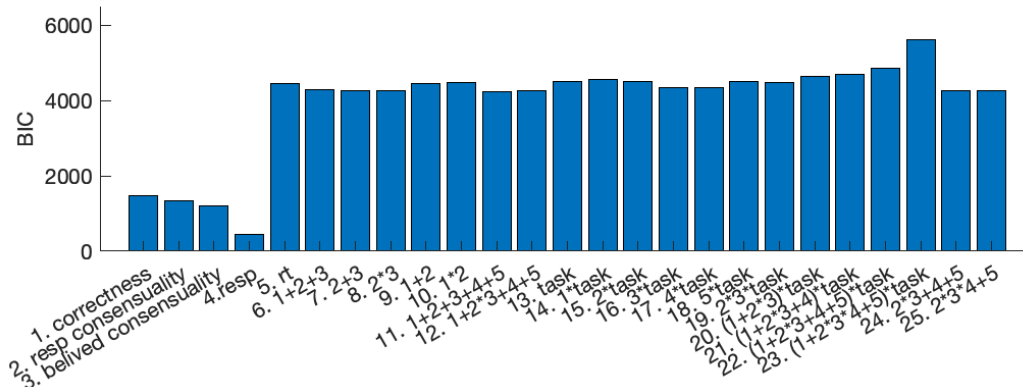


**Supplementary figure 1: Distributions of responses and reports about responses in all 4 tasks.** Four different groups of volunteers took part in respectively a 2, 3, 4 and 6 AFC tasks presenting the same paintings (c.f. sup methods). The diagrams report for each of these four tasks respectively the participants' average performance and then for all trials across participants the use of the response categories, and of the scales for confidence and believed consensuality.



**Supplementary figure 2: Mixed effects models for predictors of confidence in 2AFC.** **a.** model 9 predicting levels of confidence with actual response's correctness and response's consensuality (both encoded binarily) **b.** model 10 evaluates whether accurate consensuality predicts confidence or its prediction by consensuality **c.** model 7 predicting levels of confidence with actual response consensuality and believed response consensuality (both encoded binarily) **d.** model 8 evaluates whether accurate believed consensuality predicts confidence or its prediction by consensuality **e.** model comparison of 12 candidates to explain the factors predicting confidence levels. Lowest BIC goes to the parsimonious model 3 whereby confidence is best predicted by believed response consensuality. model 6 presenting the three central predict of the paper is presented in the main paper as figure 2a, full model 12 is presented in main paper figure 2c. Significance of predictors is indicated by \* for  $p < .05$  and \*\* for  $p < .001$ .






**Supplementary figure 3: predictors of accuracy and consensuality monitoring in task with various number of alternative choices.** **a.** model 22 presenting the effect of task as linearly increasing number of alternative (2,3,4,6) on the previously tested predictors of confidence: response correctness, response magnitude, response time, and believed and objective consensuality together with their interaction (theory of mind ability) **b-c.** simpler models of confidence presenting response consensuality, believed consensuality and response magnitude as affected by the number of alternative respectively for 2 vs 3 AFC and 2 vs 4 AFC. **d.** combined plot presenting (as if fig 2b) the prediction of believed consensuality by objective consensuality (theory of mind ability) respectively for the 2, 3, 4 and 6AFC tasks. **e.** model 24 predicting in all 4 tasks the main predictors of confidence from the full model 22 as response magnitude, response time, objective and believed consensuality together with their interaction (theory of mind ability) **f.** model 25 adding to model 24 the cues for response magnitude as interacting with the cues for theory of mind ability. **g.** model comparison for predictors of confidence across all 4 tasks as encoded with increasing number of alternative. The best model across all 4 tasks is response magnitude. Full model 22 is presented in a for effect of increasing number of alternative on predictors of confidence and 24-25 present more parsimonious models with the effect of the best predictor (response magnitude) as interacting factor to shared parameters of confidence, believed and objective consensuality. Significance of predictors is indicated by \* for  $p < .05$  and \*\* for  $p < .001$ .

### 5.3 Supplementary material

n	Market price €	Log <sub>10</sub>	size	artist	auction
1	1	0	51 x 67	Carl Weisgerber	Ketterer Kunst
2	1	0	59,5 x 42,5	Rupprecht Geiger	Ketterer Kunst
3	350	2	47x 64	Hamburg	Ketterer Kunst
4	600	2	52 x 67	Fritz Wotruba	Ketterer Kunst
5	3000	3	52 x 67	Gerhard Richter	Ketterer Kunst
6	6000	3	42,5 x 49	Bruce Nauman	Ketterer Kunst
7	20000	4	42 x 56	Ernst Ludwig Kirchner	Ketterer Kunst
8	30000	4	46 x 61.5	Julien Dupré	Ketterer Kunst
9	200000	5	50 x 60	Oskar Kokoschka	Ketterer Kunst
10	300000	5	53 x 43	Gerhard Richter	Ketterer Kunst
11	3000000	6	76.2 x 63.5	David Hockney	Christies
12	2500000	6	81.2 x 65.3	Pablo Picasso	Christies

**Supplementary Table 2: Distribution of the painting selection for their market price to span equally across log<sub>10</sub> price categories of 2,3,4 and 6 AFC tasks.**



<p>1.</p> 	<p>2.</p> 	<p>3.</p> 
<p>4.</p> 	<p>5.</p> 	<p>6.</p> 
<p>7.</p> 	<p>8.</p> 	<p>9.</p> 
<p>10.</p> 	<p>11.</p> 	<p>12.</p> 

Supplementary Table 3: Paintings used in the 4 tasks of the experiment as referred by their index as in sup table 2.



## **Bridge:**

# **From heuristics to moral values.**

This thesis argues for the role of subjective value in providing a comprehensive picture of procedural metacognition as a thermostat for decision coherence. By developing the models of both the function and the computation of metacognitive monitoring signals, we suggest that subjective value is essential to close the loop and provide a comprehensive understanding of metacognition. The second half of the thesis concerns empirical work that aims at providing new models for the computation of confidence reports in regard to subjective value.

Chapter 4 explored the cues on which confidence reports rely when the task at hand comes with limited knowledge about the decision rule. More specifically, after defining in previous chapters that subjective value is both a central input for monitoring signals and is itself relying on inference, we tested here the nature of the cues on which confidence reports rely in contexts where no explicit knowledge is yet present to define the decision rule. Our results suggested that reflective processes (i.e. both metacognition and theory of mind) rely on the same cues which guide decisions themselves both in the agent and also in the group's popular tendencies. These results, therefore, suggest a common architecture for both these reflective processes while also supporting that metacognition monitors the decision coherence with the decision rule rather than monitoring its objective accuracy.

In this investigation of metacognition, as monitoring the coherence of a decision with a decision rule, these results can be seen as one end of the spectrum: the heuristic or implicit cues as the value of items. To claim that this function of metacognition is ubiquitous, however, one must look at the other side of the

spectrum and ask: are confidence signals also influenced by subjective value when it is explicit, so much so as defined by one's conscious and voluntary choice?

When confidence reports have already been demonstrated to accurately track the coherence of decisions when concerned with the hedonic value of snacks, Chapter 5 explores whether another value domain, closer to explicit and subjective identity might also inform confidence signals: moral values.

# Chapter 5:

## Confidence monitors and predicts moral decisions.

### Abstract

When making a moral decision, we may feel more or less confident that we made the right choice. But what does this subjective confidence reflect in the moral domain? By analogy with confidence in hedonic choices, we hypothesised that in non-social context, confidence in moral choices tracks the coherence of our decisions with our moral values. To compare the two, we first asked participants to report how much they valued charities (moral domain) and snacks (hedonic domain). They then had to choose which item they preferred among pairs of charities or snacks and to report how confident they were in having chosen their favourite item. We replicated previous findings in hedonic choices and showed that participants are also able to monitor whether their choices match their own moral values. Furthermore, as in the hedonic domain, moral confidence predicted whether a decision would be consistently repeated over time. Lastly, as previously demonstrated in hedonic choices, we observe that participants whose confidence tracked better their choices' coherence with their value hierarchy also predicted better future choices' consistency. Altogether these results extending to the moral domain strengthen the evidence that metacognition both tracks both retrospective coherence and prospective consistency.

## 1. Introduction

Moral heroes are portrayed as always being sure of the rightness of their choices. In real-life however, deciding for instance to invest in an organic product rather than to give to a beggar in the street triggers a feeling of uncertainty about whether we made the right choice. Arguably, as humans, this ability to reflect upon the rightness of our own decisions defines us as rational moral agents, and supports our moral reasoning and self-evaluation in general (Paxton & Greene, 2010).

This capacity to reflect upon our own choices is studied in other domains as a form of “thinking about thinking”: a metacognitive evaluation of whether our choices comply with a given decision rule. We can for instance reflect and report whether our perceptual decisions correctly follow objective instructions (Pouget et al., 2016) or whether our choices of snacks cohere with the hierarchy of our stated hedonic preferences (De Martino et al., 2013). However, unlike perceptual instructions and hedonic preferences, moral values cannot directly be sensed or experienced. Instead, moral judgments are a complex construct of both moral intuitions (expressed as emotions and based both on innate and socio-cultural influences, Haidt, 2001; O’Neill & Petrinovich, 1998) and higher-order evaluations (including reflection and reasoning, J. D. Greene et al., 2004; J. Greene & Haidt, 2002; Paxton et al., 2012; Pizarro & Bloom, 2003; Young & Saxe, 2008). This multi-dimensional construct often associates moral judgments with uncertainty and disagreements (Bykvist, 2017; Skitka, 2010). Therefore, here we ask: Can moral values, although complex and eventually uncertain, form a clearly organised hierarchy on which one can discriminatively reflect to evaluate one’s own choice?

Confidence ratings give us an insight into how individuals reflect and subjectively evaluate their own choices. In simple value-based choices, for example a choice between two snacks (Folke et al., 2017) or everyday objects (De Martino et al., 2017), participants express higher confidence when they chose the item they value the most, and lower confidence when they chose the item they find less valuable. In other words, if an individual said that she really liked chocolate bars and just tolerated crisps, she would most likely be highly confident when choosing a

chocolate bar over the crisps, and express low confidence otherwise (De Martino et al., 2013). Therefore, even when decisions are made based on subjective values (rather than on objectively defined instructions), participants can reflect on whether their choices are coherent with their value hierarchy in a discriminate manner. This ability to reflect upon hedonic choices can also inform future choices (Boldt et al., 2019; Folke et al., 2017): when presented with the same set of items, participants are more likely to repeat their decisions if they are highly confident about it. When it comes to the moral domain, subjective uncertainty has been studied at the level of moral judgments, but whether humans have the ability to reflect upon and evaluate their own moral choices remains unexplored. It has been demonstrated for instance that people who express higher moral conviction in their moral views are more likely to invest in the corresponding causes (Skitka, 2010). In this paper, we ask instead whether participants can communicate with their confidence levels a subjective evaluation of their choices according to their own moral values, and whether they can do so as distinctively as with their hedonic values.

There are several reasons to doubt that confidence could track how our choices cohere with our moral values. First, and as already mentioned, this monitoring could be prevented because moral values would be fuzzy and uncertain due to the complexity and variability of their construct (Bykvist, 2017; J. Greene & Haidt, 2002). Moral dilemmas, in moral psychology, show that we are often conflicted and unsure of what is the right thing to do, and cannot simply solve this uncertainty by getting more information about the options we face: in a classic trolley problem, or its more recent iterations with driverless cars (Bonnefon et al., 2016; Kallioinen et al., 2019; Maxmen, 2018), we may not be sure that letting one person die to save three is the right thing to do, and looking at the options will not help us solving this tension. If moral values are particularly uncertain, there may be no clear and stable ground on which metacognition could operate allowing our confidence levels to communicate whether our choices cohere with our values. Another cue supporting the possible inherent uncertainty of moral values is brought by cognitive neuroscientists who demonstrated that altruistic decisions consistently take longer to make than self-centred ones. Choice difficulty and uncertainty are known to decrease performance

and overall confidence, and limited evidence available at the cognitive level is also considered as a limit to the metacognitive ability to monitor choices (Fleming & Daw, 2017; Fleming & Lau, 2014; Pouget et al., 2016).

A second reason to doubt that confidence could genuinely reflect moral coherence comes from self-serving biases: confidence often serves one's goals or preserve one's self-image, such as when gamblers or entrepreneurs would disregard risk and convince themselves or others of the rightness of their choice (Griffin & Tversky, 1992; Hirshleifer et al., 2012). Considering the role of moral signalling in social identity (sometimes leading to moral hypocrisy), it could be argued that confidence in moral choices would be overall tuned to fit social contexts and ambitions instead of monitoring the choices' coherence with one's actual personal values (Bogaert et al., 2008; Gal, 2015; Johnson & Chattaraman, 2020; Lönnqvist et al., 2014).

Lastly, the demonstration that the analogous self-centred versus altruistic choices are processed by different brain regions suggests that moral and hedonic values can be seen as two independent value domains (Brosch & Sander, 2013; Soutschek & Tobler, 2018; Young & Saxe, 2008). Though the debate is not settled, this difference between the way these values are encoded in the brain questions whether the same monitoring system could evaluate moral choices as well as hedonic choices. Indeed, the domain-specific hypothesis of metacognition suggests that metacognitive insight might not be ubiquitous and could vary depending on the cognitive task and types of evidence, such as between different sensory inputs or memory recall (Fitzgerald et al., 2017; Rouault et al., 2018). Although individuals are able to metacognitively reflect on their hedonic choices, this ability might therefore not generalise to moral choices.

Looking for a metacognitive ability in the moral domain is nonetheless not a lost cause. In the economics literature, the concept of "warm glow" suggests that participants experience direct pleasure when choosing what they find good for others in a somewhat similar way as they would by experiencing pleasure themselves (Andreoni, 1990). Aligning with the idea that metacognition could be domain general, pleasure could therefore serve as a common currency across value domains (Cabanac, 1992) and support a shared monitoring system for both moral



and hedonic choices types, analogous to what is demonstrated across different cognitive tasks (Baer & Odic, 2020; Mazancieux, Dinze, et al., 2020; Mazancieux, Fleming, et al., 2020; Rouault et al., 2018). Furthermore, the contextual adjustments that one sees in social contexts may affect the expressed levels of confidence (i.e, overconfidence, or bias) but not the capacity to distinguish between better and worse choices (i.e., sensitivity). More specifically, we therefore ask here whether, at least outside social pressures or audience, participants could be sensitive to the coherence between their choices and their moral values when reporting their levels of confidence, and so in a similar manner to the hedonic domain.

We were interested in two sides of metacognition: monitoring coherence and informing consistency. To test these abilities, we captured participants' hedonic values with snacks and moral values with charities. Although charities can be seen as performing a social or political lobbying role, they often collide with moral causes, and offer a good way to investigate moral preferences (Maoz et al., 2019; Nilsson et al., 2016). We started by recording participants' hierarchy of values in both hedonic and moral by asking them to rate their likings of sets of snacks and charities (following a standard procedure, e.g. Brosch & Sander, 2013; Colas, 2017; Colosio et al., 2017; Harbaugh et al., 2007; Hare et al., 2010; Maoz et al., 2019; Moll et al., 2006; Sepulveda et al., 2020; Tarantola et al., 2017). Participants then saw pairs of snacks or charities and had to choose which of the two items they preferred before reporting how confident they were in having chosen their favourite item. We hypothesized that subjective confidence levels would discriminately track whether choices were coherent with the participants' hierarchy of moral values. Furthermore, if such moral metacognitive monitoring exists, we expected that confidence levels could also shape future behaviour, by predicting whether a choice would be repeated or not. Following previous findings in the hedonic and economic domain (Boldt et al., 2019; Folke et al., 2017), we then expected to find confidence as a predictor of choice consistency in the moral domain.

Our results replicated existing findings in the hedonic domain and extend them to the moral domain by showing that participants' confidence both tracked the

coherence of their choices and guides future choices. In both value domains, individuals with greater insight into the coherence of their choices also had more informative confidence levels in predicting their future behavioural consistency.

## **2. Methods**

### **2.1. Participants**

A total of 43 participants fluent in English were recruited to take part in the computer-based experiment at LMU's psychology laboratory. To test metacognitive ability in charity choices, we used a similar sample size to previous studies which demonstrated metacognitive ability in snack choices (De Martino et al., 2013; Folke et al., 2017). All participants signed a consent form and were compensated 9 euros per hour with a 1 euro bonus if they performed above 85% in the attention checks and comprehension questions at the end of the experiment. We rejected 6 participants whose choices' coherence score (according to reported likings) was out of the 60-95% range, in either value domain. The remaining 37 participants were included in the analyses (17 females, aged: 20-43). The study was approved by the University of London Research Ethics Committee (Project Number: SASREC\_1819-313A).

### **2.2. Stimuli**

A total of 16 snacks were selected for their differences and similarities along 4 axes: sweet/savoury, healthy/rich, single item/pack of items and rare/popular snack. For charities, the selection was based on Maoz et al., 2019 's bank of stimuli from which 8 consensual charities and 4 pairs of controversial charities were selected. This selection was made to allow similarities and differences across 6 moral causes: conservation (of environment and biodiversity), (human) rights, health (and research), support (to people in need), (social) inequalities and culture (and education). A pilot study on 31 participants was used to best homogenize the value distribution between snacks and charities across participants (Fig. S1a-b). High-definition pictures of the snacks and of the charities' names and logo were downloaded from the internet. Finally, to normalize the information available for all

stimuli, we applied a standard format of 400x400 pixels and added a few words to describe the snacks and the charities. The task was coded in JavaScript (JSpsych.org).

### **2.3. Procedure**

After reading general instructions and consent forms on their computer screen, participants saw a snapshot of all the options used in the experiment (8 random at a time, for 20 seconds) and were asked to familiarise themselves with the sets of items. This first glance aimed at helping participants to use the value scale more finely by having a prior knowledge of the overall range of snacks and charities that would be used.

*Value report.* After a 500ms fixation cross, participants saw either a snack or a charity and rated how much they would like to respectively *obtain* this snack or *support* this charity on a continuous scale from *really dislike* to *really like* (quantified as -10 to 10, step .5, Fig. S1a-b, Lebreton et al., 2015). This continuous scale aimed at preventing participants from thinking in a quantitative manner so that the following choice task would be performed genuinely and not by memorising the explicitly reported values in this part. Furthermore, for both snacks and charities, this personal report of liking aimed at eliciting the participants' subjective values of each item instead its more objectively defined price, popularity or rightness. After each evaluation, participants rated how certain they were of this reported value, on a continuous scale from *not at all* to *absolutely* certain (Fig. 1a). As previously reported in snack choices (De Martino et al., 2013), post hoc analyses did not reveal a significant link between this rating of value certainty and choice confidence (Fig. S3d), therefore this measure is not further mentioned in the present paper. This value report was done in 2 blocks of either 16 snacks or 16 charities, and both the order of the blocks and of the items were randomised.

*Choice.* All combinations of two snacks or two charities (N=120 each) were then presented twice in reversed lateral position (480 trials), across 8 blocks which were presented in random order and contained a series of 60 random choices between pairs of items of the same domain. In each trial, participants saw a 500ms fixation cross, and then either two snacks or two charities. They were asked to select the item they would prefer to respectively *obtain* or *support*. After each decision, participants

were asked how confident they were that they chose their preferred item from *not at all* to *absolutely* confident (quantified for analysis as 0 to 10 step .1). No time pressure was applied in the entire task but decisions that took over 3 STD beyond the subject's average response time in the domain were discarded ( $5.1 \pm 1.5\%$  of trials). All parameters were z-scored independently within participants and domains to favour their fair comparisons in mixed models and other analyses. A five minutes demo of the task is available at:

<https://www.cvbe-experiments.com/oa/MoralConfidenceDemo/>

### **3. Results**

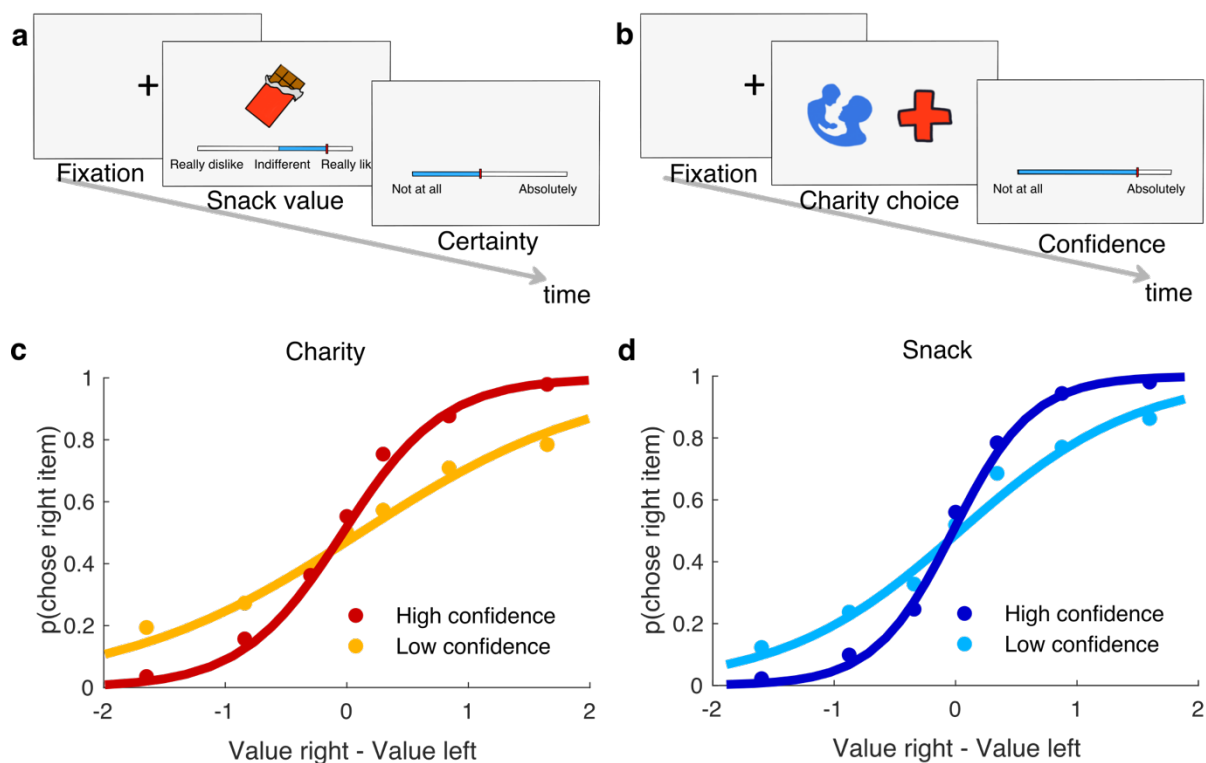
#### ***3.1. Choice and confidence's sensitivity to coherence.***

Following a standard procedure in behavioural economics, we first asked participants to separately rate how much they valued each item (Becker et al., 1964). We then asked participants to choose among pairs of charities and snacks. We first tested whether participants' choices amongst charities were coherent with their moral values, by which we mean in this paper that participants chose the item for which they expressed the highest value. Following our expectations, we found that participants' choices among charities were indeed coherent, hence demonstrating that their moral decisions could be predicted by their explicitly reported moral values (Fig. S1d). More specifically, this choice coherence was predicted by choice difficulty, namely, participants were more likely to choose the charity they rated as more valuable when the difference in value (DV) between both charities was high (easy choices, Sugrue et al., 2005, logistic regression:  $\beta = 2.79$ ,  $se = 0.27$ ,  $z = 10.42$ ,  $p < 0.001$ ). Replicating previous findings, participants also exhibited a coherent behaviour in their choices of snacks, therefore allowing us to study the relationship between confidence and this choice coherence in both domains ( $\beta = 3.40$ ,  $se = 0.28$ ,  $z = 11.94$ ,  $p < 0.001$ ).

We then tested whether participants' confidence levels captured whether their moral choices were coherent with their subjective values, that is, we asked: were decisions reported with high confidence more likely have chosen the highest value item than decisions reported with low confidence? We found that confidence levels discriminated coherent from incoherent moral choices (Fig. 1c, S2e paired t-test:  $t =$

6.63,  $p < 0.001$ ). In other words, in choices with high confidence, participants chose more frequently the charity they rated as more valuable than in low confidence choices. This sensitivity of confidence to coherence was confirmed by the interaction between these parameters making the best model to predict the moral choices made by participants (Fig. S2d:  $\beta = 0.85$ ,  $se = 0.08$ ,  $z = 10.12$ ,  $p < 0.001$ , Fig. S2a Bayesian Information Criterion (BIC) relative to model 3 DVxConfidence: 7394; model 1 DV: 7631; model 2 DV+Confidence: 7650). Replicating previous findings, we found that this sensitivity of confidence also applied in choices among snacks (Fig. S2d:  $\beta = 1.20$ ,  $se = 0.12$ ,  $z = 9.80$ ,  $p < 0.001$ , Fig. S2b BIC relative to model 3 DVxConfidence: 6425; model 1 DV: 6781; model 2 DV+Confidence: 6803; Fig. S2e paired t-test:  $t = -6.17$ ,  $p < 0.001$ , De Martino et al., 2013; Folke et al., 2017).

Altogether, our results demonstrate that confidence levels are reliable explicit markers of the coherence between one's choice and one's personal moral values. Namely, participants successfully manage to evaluate how their choices reflect their personal moral values and have explicit access to this appraisal. Our replication of this sensitivity of confidence in snack choices highlights that healthy individuals are explicitly aware of their choices' coherence with their personal values, and so both when these concern their personal interests in food and their views for others' welfare.



**Figure 1: Choice and confidence sensitivities a.** Participants first rated how much they valued each charity and snack and their certainty in this rating on continuous scales. **b.** They were then presented with pairs of either charities or snacks and asked to choose the item they would prefer to respectively *support* or *obtain* and report their confidence in having chosen their favourite item. **c-d.** Probability of choosing item on the right side given its relative value (DV: difference in value) to the item on the left (logistic fit for all participants, respectively for charity and snack choices). This choice sensitivity was modulated by confidence levels, this difference between the fit for low and high confidence levels (subjective median split) is a proxy for subjective metacognitive accuracy (for mixed model see Fig. S2d).

### **3.2. Confidence predicts choice consistency.**

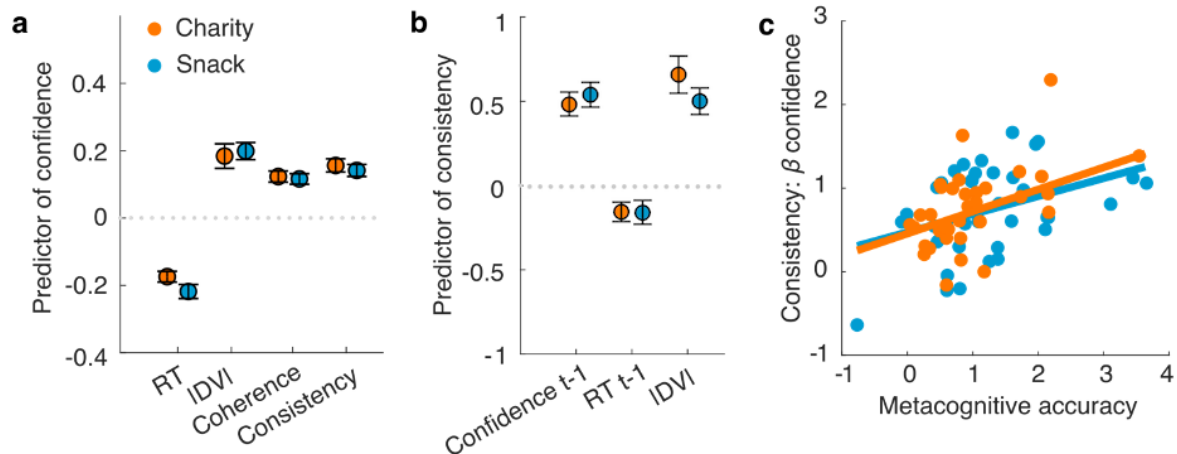
In economics, a choice behaviour is rational if it follows a series of norms such as coherence with one's order of preferences, consistency over time, or transitivity among choices. Since confidence was sensitive to the choice's coherence with one's values, we hypothesized that confidence could also monitor whether a choice was consistent over time, if the same set of items was presented several times. For both charity and snack choices, we presented participants with each pair of items twice (with counter-balanced item position) enabling us to investigate whether decisions which were consistently repeated were reported with higher confidence than the inconsistent decisions. A mixed model (Fig. 2a) demonstrated indeed that choice consistency predicted the level of confidence in the choice (Charity:  $\beta=0.15$ ,  $se=0.02$ ,  $z=7.92$ ,  $p<0.001$ , Snack:  $\beta =0.14$ ,  $se=0.02$ ,  $z=8.43$ ,  $p<0.001$ ). Furthermore, the prediction of confidence by choice consistency seemed to be at least as strong as for choice coherence (from best model (7) Fig. S3c: Charity consistency:  $\beta =0.16$ ,  $se=0.02$ ,  $z=9.53$ ,  $p<0.001$ , Charity coherence:  $\beta=0.10$ ,  $se=0.01$ ,  $z=8.42$ ,  $p<0.001$ , Snack consistency:  $\beta=0.14$ ,  $se=0.02$ ,  $z=8.81$ ,  $p<0.001$ , Snack coherence:  $\beta=0.10$ ,  $se=0.02$ ,  $z=6.12$ ,  $p<0.001$ ). This effect was confirmed by a model comparison between the effect of both these norms of rationality on confidence (Fig. S3c, BIC scores: consistency model 2= 46734; coherence model 1= 46843). Therefore, for moral and hedonic value domains, our results highlight that confidence tracks, further than

simply the coherence of one's choice with their preferences, also choice consistency over time.

Building on this strong effect of choice consistency on confidence level, a second analysis focused on the reverse relation: how do confidence levels, amongst other factors, predict the consistency of future choices (Flavell, 1979; Folke et al., 2017; Kluwe, 1982; Nelson & Narens, 1990)? To do so, we investigated whether the confidence level in the first encounter of the choice (t-1) could predict whether the same decision would be made again when the same items pairing is presented once more (t). Among other candidate factors, a mixed model investigated the effect of confidence at the previous encounter of the items pair in predicting whether the second decision would be consistent with the first one (Fig. 2b). As could be expected, choice difficulty ( $|DV|$ ) best predictor of choice coherence from Fig. 1c,d), but also confidence at t-1 were strong predictors of choice consistency at t (both factors controlled within same model Fig. 3b Charity  $|DV|$ :  $\beta=0.60$ ,  $se=0.10$ ,  $z=5.90$ ,  $p<0.001$ , Snack  $|DV|$ :  $\beta=0.50$ ,  $se=0.08$ ,  $z=5.90$ ,  $p<0.001$ , Charity confidence t-1:  $\beta=0.51$ ,  $se=0.06$ ,  $z=8.85$ ,  $p<0.001$ , Snack confidence t-1:  $\beta=0.54$ ,  $se=0.07$ ,  $z=7.38$ ,  $p<0.001$ ). A model comparison also suggested that confidence levels in the previous choice (t-1) explained at least as much about choice consistency than choice difficulty itself (Fig S4a-b BIC scores for Charity model 2 confidence=3218, model 1  $|DV|=3278$ , for Snack model 2 confidence=2771, model 1  $|DV|=2878$ ). In other words, regardless of whether a choice between charities was obvious or difficult to participants, how confident they were the first time they made this choice was at least as likely to predict whether they would make the same decision again.

Furthermore, response time, often considered as an implicit marker of uncertainty in one's decision (Faivre et al., 2018; Folke et al., 2017) was a small predictor of choice consistency both in snack and charity choices (Fig. 2b, Charity:  $\beta=-0.07$ ,  $se=0.06$ ,  $z=-1.23$ ,  $p=0.22$ , Snack:  $\beta=-0.26$ ,  $se=0.05$ ,  $z=-5.66$ ,  $p<0.001$ , Fig S1e-f, response time (implicit choice uncertainty) predicts explicit confidence: paired t-test Charity:  $t=8.20$ ,  $p<0.001$ , Snack:  $t=10.80$ ,  $p<0.001$ , and coherence: paired t-test Charity:  $t=14.11$ ,  $p<0.001$ , Snack:  $t=15.62$ ,  $p<0.001$ ). These results suggest that, independently of the average longer time and associated speed accuracy trade off in charity choices, participants might be able to cue their confidence levels on a domain relative

response time as an implicit maker of choice reliability. While both implicit and explicit markers of choice reliability predict consistency, explicit reports of confidence were stronger predictors. Altogether, we found that confidence in a moral choice, further than reflecting whether the choice is coherent with one's personal values, is also predictive of repeated decisions in the future.



**Figure 2: Confidence predicts choice consistency** **a.** Predictors of confidence (mixed model 6 from model comparison Fig. S3c, full model Fig. S3d) with respectively response time (RT), difference in items' values ( $|DV|$ ), choosing high-value item (Coherence), repeating the choice (Consistency) **b.** Fixed effects predicting the choice consistency at  $t$  with the previous choice amongst this pair of items at  $t-1$  (mixed model 4 from model comparison Fig. S4c, full model Fig. S4d). **c.** Linear regression between participants' metacognitive accuracy in monitoring choice coherence (from the best model (7): Fig. S2d, similar to simplified Fig. 1 c-d) and participant's predictor of choice consistency by confidence at  $t-1$  ( $\beta$ Confidence  $t-1$  from the best model of consistency Fig. S4d, similar to panel b).

### 3.3. Metacognition monitors and predicts.

Lastly, we tested whether, at the individual level, participants whose confidence was more informative about the value of their choices (metacognitive accuracy) also had a more predictable behaviour based on these confidence levels. In other words, we investigated the link between metacognitive insight and a possible metacognitive behavioural prediction by studying their relationship within individuals. Our results revealed that in the moral domain, participants with higher metacognitive insight in their choices' coherence also were those whose confidence predicted best their



future choices' consistency (Fig. 2c  $\beta=0.26$ ,  $se=0.08$ ,  $z=3.22$ ,  $p<0.01$ ). While this effect was previously demonstrated in snack choices (Folke et al., 2017), in our experiment these results were only approaching significance without reaching ( $\beta=0.21$ ,  $se=0.10$ ,  $z=2.21$ ,  $p=0.05$ ). These results overall suggest that, in both domains, individuals with greater metacognitive insight were more likely to repeat a choice if they were highly confident the first time they made their choice and to go for the alternative item otherwise, whereas, for participants with lower metacognitive insight, confidence did not have such a strong predictive power. Our results therefore suggest that participants with greater ability to reflect on how valuable their choices were to them could also partly rely on this signal to inform future decision making processes. Nonetheless other empirical design would need to confirm a role of causality between this monitoring and guiding.

#### **4. Discussion**

In our study, we present a novel approach to evaluate whether participants are able to subjectively monitor how their choices cohere with their moral values. While the ability to consciously monitor whether one's choice is coherent with one's hedonic values was previously demonstrated (De Martino et al., 2013; Folke et al., 2017), here we tested this metacognitive ability in the moral domain. If economists and psychologists can describe confidence in moral choices as self-serving or motivational, especially in social contexts, we investigated whether confidence could instead use subjective moral values to evaluate one's choices. Our results demonstrated this metacognitive ability to monitor choices' coherence and also to predict future choices' consistency in the moral domain while replicating these findings in the hedonic one. These results highlight that, in both choices concerning either oneself or others, metacognition allows conscious monitoring of choices and might contribute to guiding future choices with explicit confidence levels.

We challenged the metacognitive ability to track choice coherence in the moral domain for the reasons that values and decisions in this domain are often associated with uncertainty and complexity (Bykvist, 2017; J. Greene & Haidt, 2002; Paxton et al., 2012). At this cognitive level however, our experiment did not reveal a difference

of uncertainty between the ratings of charities and snacks (Fig. S1c). Although uncertainty and response time are commonly linked in perceptual choices, our experiment replicated the difference in response time found between both domains whereby moral choices take consistently longer than hedonic choices. (Crockett et al., 2015; Krajbich et al., 2015; Moll et al., 2006). These findings highlight that the decision processes between both value domains appear to rely on different mechanisms, and supports the findings that more reflective cognitive processes might be at play to make moral decisions than hedonic ones (J. D. Greene et al., 2004; J. Greene & Haidt, 2002; Paxton et al., 2012; Pizarro & Bloom, 2003; Schenk, 2006; Young & Saxe, 2008). Therefore, in our study, the moral domain was not necessarily associated with more uncertainty in subjective values but rather with a different and eventually more demanding decision process.

Despite these differences between domains at the cognitive level, we investigated whether metacognition could monitor how choices related to moral values as well as it does with hedonic values. If it is agreed that the cognitive and metacognitive processes use at least partly the same information, how such cognitive links between both levels relate their respective behavioural performances together is still unclear (Fleming & Daw, 2017; Fleming & Lau, 2014). Our results demonstrated comparable metacognitive abilities in both value domains, therefore suggesting that confidence could relate to different types of value in the same way. This similar metacognitive access to different decision processes was proposed to be mediated by supra-modal cues (e.g. response time, Faivre et al., 2018) or processes (e.g. working memory, Shea & Frith, 2019). Ultimately, a comparative evaluation of choices across tasks could serve as a common currency to guide behaviour in a multi-dimensional environment (De Gardelle et al., 2016). The metacognitive access to moral decision-making process could be investigated in a clinical population with altered altruism such as the psychopathic population (Abu-Akel et al., 2015; Bo et al., 2014). While for instance blind-sight studies have proved metacognition to access unconscious information (not available at the cognitive level, Ko & Lau, 2012), such residual metacognitive access could suggest non-null metacognitive monitoring even with an impaired moral system. Lastly, the demonstration that participants tend to act

against their moral values in social context could predict to reduce the metacognitive ability to monitor how choices cohere with personal values by changing the individual's goal in this setting (Griffin & Tversky, 1992; Hirshleifer et al., 2012).

While we demonstrated overall comparable metacognitive abilities to monitor choices in both hedonic and moral domain, our design's lack of control for value hierarchies (Fig. S1a-b) and coherence levels (Fig. S1d) across domains limited the extent to which we could compare these confidence monitoring processes. By selecting which snack pairs were presented to participants based on their subjective value, Folke *et al.* (2017) partly controlled the cognitive process monitored by confidence, bringing the value-based decision paradigm closer to the perceptual literature. By controlling for choice difficulty, the equal amount of evidence in both value domains would bring a common cognitive ground on which to quantify and compare metacognitive abilities between domains. On the one hand, the complexity of moral judgements could make for an unclear hierarchy of values on which to found metacognitive monitoring, therefore predicting lower metacognitive ability in the moral than hedonic domain. On the other hand, if complex moral values are uniquely encoded by higher-order cognitive processes (e.g. reflective system) adjacent to metacognitive monitoring, one could expect that metacognitive insight might be greater in the moral than the hedonic domain (J. D. Greene et al., 2004; Paxton et al., 2012; Pizarro & Bloom, 2003; Schenk, 2006). Additionally, studying the link between metacognitive abilities across value domains and within participants could shed light on the metacognitive system by revealing whether there could be a unique overarching metacognitive system for all value-based choices as suggested by the domain general hypothesis of metacognition (Morales et al., 2018; Rouault et al., 2018). To best account for these conflicting predictions, further research should ideally also investigate the contribution of the systematic factors involved this metacognitive monitoring (Shekhar & Rahnev, 2020).

Independently from the ability of confidence to monitor the coherence of choices with subjective values (i.e. sensitivity), controlling for comparable value-

based choices across domains could allow comparing the calibration of confidence (i.e. confidence bias: over- or under-confidence, Baranski & Petrusic, 1994; Fleming & Lau, 2014). For instance, self-serving bias could predict participants to report overall higher confidence in the moral domain than in the hedonic one such as to influence others (Bogaert et al., 2008; Gal, 2015; Johnson & Chattaraman, 2020; Lönnqvist et al., 2014). Alternatively, the complex construct of moral values and slower decision-making process could instead predict participants to be relatively under-confident in this domain (Crockett et al., 2015; Krajbich et al., 2015; Moll et al., 2006; Patel et al., 2012).

Our secondary hypothesis concerned the function of confidence in informing future behaviour. Aligning with previous findings, we found that, in both value domains, explicit reports of confidence predicted choice consistency over time. Interestingly, our results also replicated that these explicit monitoring signals (i.d. confidence) were more reliable predictors of future behaviour than implicit markers of uncertainty such as reaction time, and so especially in the moral domain (Folke et al., 2017). The extent to which confidence informs future choices could be more finely tested in a new experimental design by repeating choices more than twice. Furthermore, besides choice consistency, confidence could be tested for its potential guidance of other types of commitments (e.g. monetary investment, Soutschek & Tobler, 2020) or behavioural optimisation (e.g. choice transitivity, Folke et al., 2017, or confidence calibration, Rouault et al., 2019).

Lastly, we investigated within each individual the role of metacognition as a possible gateway between behavioural monitoring and control. Replicating this finding in the hedonic domain, we demonstrated that individuals whose confidence monitors best choice coherence also informs better choice consistency. To test further the possible role of metacognition in behavioural guidance, its link to deficient executive control could be studied in the clinical population such as obsessive-compulsive disorder. Indeed previous findings suggested that this population's perceptual metacognitive ability was not affected but only the guiding role of confidence for future choices was impaired (Vaghi et al., 2017). Studying in this population how confidence bridges

monitoring and control of value-based choices (which are central to impulsive and addictive behaviours) would refine the link between these metacognitive functions. By suggesting a link between metacognitive sensitivity and behavioural flexibility in value based-choice, our present study supports the existing consideration that metacognitive training (also transferable across tasks Carpenter et al., 2019) appears as a promising area of therapy for the population with psychological disorders such as with reduced executive control (Bang et al., 2020; Bhome et al., 2019; Faivre et al., 2019; Heyes et al., 2020; Lysaker et al., 2014; Vaghi et al., 2017).

### **5. Conclusion**

In the present study, we demonstrated the ability of healthy participants to reflect upon their moral choices by explicitly reporting through their confidence levels whether their choices were coherent with their own moral values. Additionally, this reflective ability also predicted how much a participant's confidence levels predicted her future choice consistency. Altogether our findings paint a picture of confidence as a gateway between the monitoring of past choices and the information of future choices in both the hedonic and moral domains.

### **Acknowledgments**

We wish to thank the Graduate School for Systemic Neurosciences LMU for supporting this doctoral project, Lucas Battich for his technical guidance in designing the study and Justin Sulik for his insight in the data analysis. We also thank the insightful audience for their feedback on the experiment at the Workshop on Consciousness, Agency and Metacognition, the Summer School on Social Cognition and at the Berlin-Munich Seminar for Behavioural Economics.

### **Funding**

OD was supported by a grant from the Excellence Initiative at LMU, and a grant from the NOMIS foundation (acronym DISE).

### **Data availability**

The data presented in this article can be found on OSF at:

## References

- Abu-Akel, A., Heinke, D., Gillespie, S. M., Mitchell, I. J., & Bo, S. (2015). Metacognitive impairments in schizophrenia are arrested at extreme levels of psychopathy: The cut-off effect. *Journal of Abnormal Psychology, 124*(4), 1102–1109. <https://doi.org/10.1037/abn0000096>
- Andreoni, J. (1990). Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving. *The Economic Journal, 100*(401), 464. <https://doi.org/10.2307/2234133>
- Baer, C., & Odic, D. (2020). Children flexibly compare their confidence within and across perceptual domains. *Developmental Psychology, 56*(11), 2095–2101. <https://doi.org/10.1037/dev0001100>
- Bang, D., Ershadmanesh, S., Nili, H., & Fleming, S. M. (2020). Private-public mappings in human prefrontal cortex. *BioRxiv*, 2020.02.21.954305. <https://doi.org/10.1101/2020.02.21.954305>
- Baranski, J. V., & Petrusic, W. M. (1994). The calibration and resolution of confidence in perceptual judgments. *Perception & Psychophysics, 55*(4), 412–428. <https://doi.org/10.3758/BF03205299>
- Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single response sequential method. *Behavioral Science, 1*, 226–232. <https://doi.org/10.1002/BS.3830090304>
- Bhome, R., McWilliams, A., Huntley, J. D., Fleming, S. M., & Howard, R. J. (2019). Metacognition in functional cognitive disorder- a potential mechanism and treatment target. *Cognitive Neuropsychiatry, 24*(5), 311–321. <https://doi.org/10.1080/13546805.2019.1651708>
- Bo, S., Abu-Akel, A., & Kongerslev, M. (2014). Metacognition as a Framework to Understanding the Occurrence of Aggression and Violence in Patients with Schizophrenia. *Social Cognition and Metacognition in Schizophrenia: Psychopathology and Treatment Approaches*, 137–149. <https://doi.org/10.1016/B978-0-12-405172-0.00008-9>
- Bogaert, S., Boone, C., & Declerck, C. (2008). Social value orientation and

- cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, 47(3), 453–480.  
<https://doi.org/10.1348/014466607X244970>
- Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. *Neuroscience of Consciousness*, 2019(1), 1–18. <https://doi.org/10.1093/nc/niz004>
- Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293), 1573–1576.  
<https://doi.org/10.1126/science.aaf2654>
- Brosch, T., & Sander, D. (2013). Neurocognitive mechanisms underlying value-based decision-making: From core values to economic value. *Frontiers in Human Neuroscience*, 7(JUL), 1–8. <https://doi.org/10.3389/fnhum.2013.00398>
- Bykvist, K. (2017). Moral uncertainty. *Philosophy Compass*, 12(3), 1–8.  
<https://doi.org/10.1111/phc3.12408>
- Cabanac, M. (1992). Pleasure: the common currency. *Journal of Theoretical Biology*, 155(2), 173–200. [https://doi.org/10.1016/S0022-5193\(05\)80594-6](https://doi.org/10.1016/S0022-5193(05)80594-6)
- Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. *Journal of Experimental Psychology: General*, 148(1), 51–64.  
<https://doi.org/10.1037/xge0000505>
- Colas, J. T. (2017). Value-based decision making via sequential sampling with hierarchical competition and attentional modulation. In *PLoS ONE* (Vol. 12, Issue 10). <https://doi.org/10.1371/journal.pone.0186822>
- Colosio, M., Shestakova, A., Nikulin, V. V., Blagovechtchenski, E., & Klucharev, V. (2017). Neural mechanisms of cognitive dissonance (Revised): An EEG study. *Journal of Neuroscience*, 37(20), 5074–5083.  
<https://doi.org/10.1523/JNEUROSCI.3209-16.2017>
- Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2015). Erratum: Harm to others outweighs harm to self in moral decision making (Proc Natl Acad Sci USA (2014) 111 (17320-17325) DOI: 10.1073/pnas.1408988111). *Proceedings of the National Academy of Sciences of the United States of America*, 112(4), e381.

- <https://doi.org/10.1073/pnas.1424572112>
- De Gardelle, V., Le Corre, F., & Mamassian, P. (2016). Confidence as a common currency between vision and audition. *PLoS ONE*, *11*(1).  
<https://doi.org/10.1371/journal.pone.0147901>
- De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. *Journal of Neuroscience*, *37*(25), 6066–6074.  
<https://doi.org/10.1523/JNEUROSCI.3880-16.2017>
- De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.  
<https://doi.org/10.1038/nn.3279>
- Faivre, N., Filevich, E., Solovey, G., Kühn, S., & Blanke, O. (2018). Behavioural, modeling, and electrophysiological evidence for domain-generality in human metacognition. *The Journal of Neuroscience*, *38*(2), 263–277.  
<https://doi.org/10.1523/JNEUROSCI.0322-17.2017>
- Faivre, N., Pereira, M., Gardelle, V. De, & Vergnaud, J. (2019). *Confidence in perceptual decision-making is preserved in schizophrenia*. *MedRxiv*.  
<https://doi.org/10.1101/2019.12.15.19014969>
- Fitzgerald, L. M., Arvaneh, M., & Dockree, P. M. (2017). Domain-specific and domain-general processes underlying metacognitive judgments. *Consciousness and Cognition*, *49*, 264–277.  
<https://doi.org/10.1016/j.concog.2017.01.011>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911.  
<https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. <https://doi.org/10.1037/rev0000045>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(JULY). <https://doi.org/10.3389/fnhum.2014.00443>
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature*



- Human Behaviour*, 1(1), 17–19. <https://doi.org/10.1038/s41562-016-0002>
- Gal, D. (2015). Identity-Signaling Behavior. In M. I. Norton, D. D. Rucker, & C. Lambertson (Eds.), *Cambridge handbooks in psychology. The Cambridge handbook of consumer psychology* (pp. 257–281). Cambridge University Press. <https://doi.org/10.1017/CBO9781107706552.010>
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400. <https://doi.org/10.1016/j.neuron.2004.09.027>
- Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), 517–523. [https://doi.org/10.1016/S1364-6613\(02\)02011-9](https://doi.org/10.1016/S1364-6613(02)02011-9)
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*. [https://doi.org/10.1016/0010-0285\(92\)90013-R](https://doi.org/10.1016/0010-0285(92)90013-R)
- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834. <https://doi.org/10.1037/0033-295X.108.4.814>
- Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations. *Science*, 6, 1622–1625. <https://doi.org/10.1126/science.1140738>
- Hare, T. A., Camerer, C. F., Knoepfle, D. T., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *Journal of Neuroscience*, 30(2), 583–590. <https://doi.org/10.1523/JNEUROSCI.4089-09.2010>
- Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences*, 24(5), 349–362. <https://doi.org/10.1016/j.tics.2020.02.007>
- Hirshleifer, D., Low, A., & Teoh, S. H. (2012). Are Overconfident CEOs Better Innovators. *Journal of Finance*, 67(4), 1457–1498. <https://doi.org/10.1111/j.1540-6261.2012.01753.x>
- Johnson, O., & Chattaraman, V. (2020). Signaling socially responsible consumption

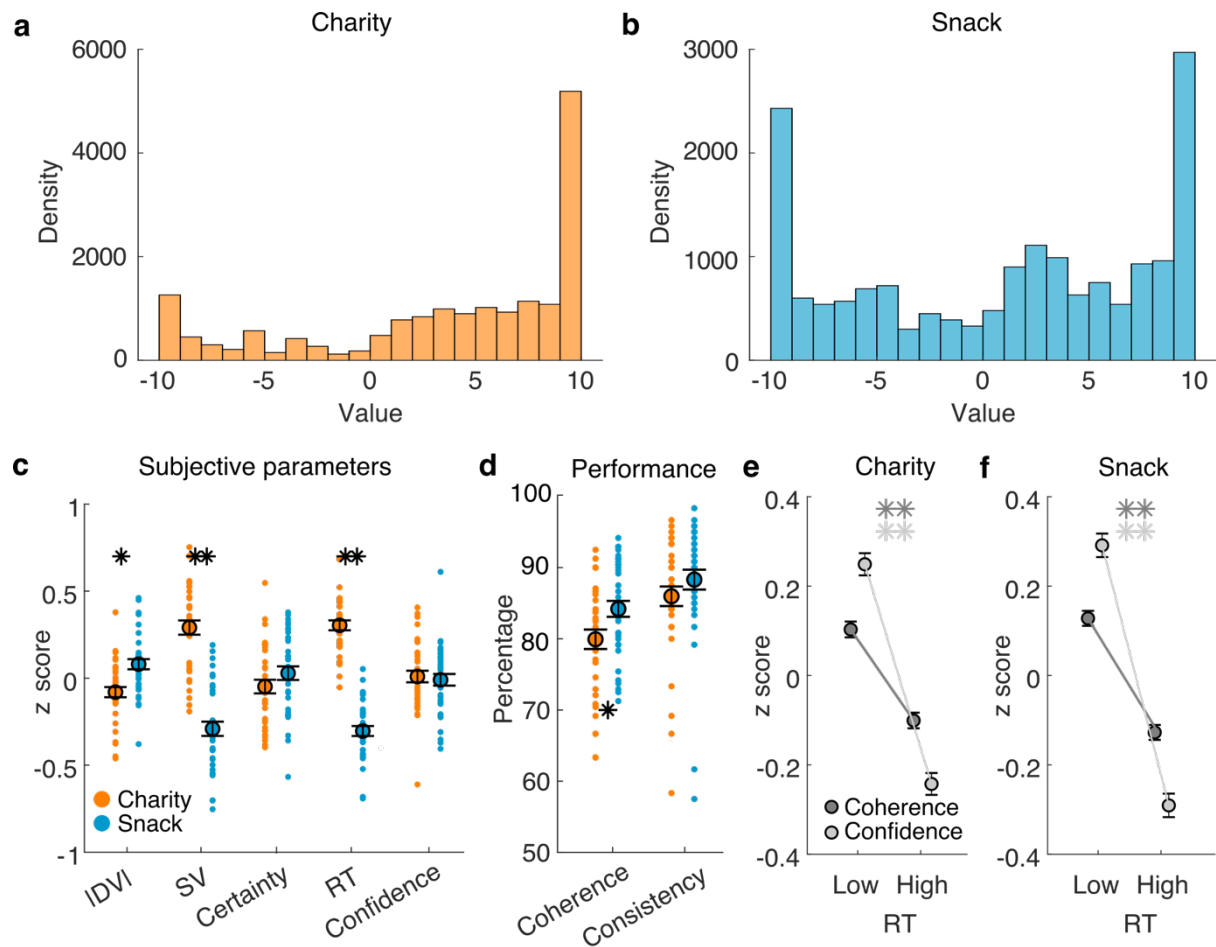
- among millennials: an identity-based perspective. *Social Responsibility Journal*, November. <https://doi.org/10.1108/SRJ-02-2019-0074>
- Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A., & König, P. (2019). Moral Judgements on the Actions of Self-Driving Cars and Human Drivers in Dilemma Situations From Different Perspectives. *Frontiers in Psychology*, *10*, 1–15. <https://doi.org/10.3389/fpsyg.2019.02415>
- Kluwe, R. H. (1982). Cognitive Knowledge and Executive Control: Metacognition. *Animal Mind — Human Mind*, 201–224. [https://doi.org/10.1007/978-3-642-68469-2\\_12](https://doi.org/10.1007/978-3-642-68469-2_12)
- Ko, Y., & Lau, H. (2012). A detection theoretic explanation of blindsight suggests a link between conscious perception and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1401–1411. <https://doi.org/10.1098/rstb.2011.0380>
- Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A Common Mechanism Underlying Food Choice and Social Decisions. *PLoS Computational Biology*, *11*(10), 1–24. <https://doi.org/10.1371/journal.pcbi.1004371>
- Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. *Nature Neuroscience*, *18*(8), 1159–1167. <https://doi.org/10.1038/nn.4064>
- Lönqvist, J. E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: Impression management or self-deception? *Journal of Experimental Social Psychology*, *55*, 53–62. <https://doi.org/10.1016/j.jesp.2014.06.004>
- Lysaker, P. H., Hillis, J., Leonhardt, B. L., Kukla, M., & Buck, K. D. (2014). Metacognition in Schizophrenia Spectrum Disorders: Methods of Assessment and Associations with Neurocognition, Symptoms, Cognitive Style and Function. *Isr J Psychiatry Relat Sci.*, *51*(1), 54–62. <https://doi.org/10.1016/B978-0-12-405172-0.00006-5>
- Maoz, U., Yaffe, G., Koch, C., & Mudrik, L. (2019). Neural precursors of decisions that matter—an ERP study of deliberate and arbitrary choice. *ELife*, *8*, 1–23. <https://doi.org/10.7554/eLife.39787>
- Maxmen, A. (2018). Self-driving car dilemmas reveal that moral choices are not

- universal. *Nature*, 562(7728), 469–470. <https://doi.org/10.1038/d41586-018-07135-0>
- Mazancieux, A., Dinze, C., Souchay, C., & Moulin, C. J. A. (2020). Metacognitive domain specificity in feeling-of-knowing but not retrospective confidence. *Neuroscience of Consciousness*, 2020(1), 1–11. <https://doi.org/10.1093/nc/niaa001>
- Mazancieux, A., Fleming, S. M., Souchay, C., & Moulin, C. J. A. (2020). Is there a G factor for metacognition? Correlations in retrospective metacognitive sensitivity across tasks. *Journal of Experimental Psychology: General*, 149(9), 1788–1799. <https://doi.org/10.1037/xge0000746>
- Moll, J., Krueger, F., Zahn, R., Pardini, M., De Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America*, 103(42), 15623–15628. <https://doi.org/10.1073/pnas.0604475103>
- Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, 38(14), 3534–3546. <https://doi.org/10.1523/JNEUROSCI.2360-17.2018>
- Nelson, T. O. (1990). Metamemory: A Theoretical Framework and New Findings. *Psychology of Learning and Motivation - Advances in Research and Theory*, 26(C), 125–173. [https://doi.org/10.1016/S0079-7421\(08\)60053-5](https://doi.org/10.1016/S0079-7421(08)60053-5)
- Nilsson, A., Erlandsson, A., & Västfjäll, D. (2016). The congruency between moral foundations and intentions to donate, self-reported donations, and actual donations to charity. *Journal of Research in Personality*, 65, 22–29. <https://doi.org/10.1016/j.jrp.2016.07.001>
- O'Neill, P., & Petrinovich, L. (1998). A Preliminary Cross-Cultural Study of Moral Intuitions. *Evolution and Human Behavior*, 19(6), 349–367. [https://doi.org/10.1016/S1090-5138\(98\)00030-0](https://doi.org/10.1016/S1090-5138(98)00030-0)
- Patel, D., Fleming, S. M., & Kilner, J. M. (2012). Inferring subjective states through the observation of actions. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748), 4853–4860. <https://doi.org/10.1098/rspb.2012.1847>
- Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics*

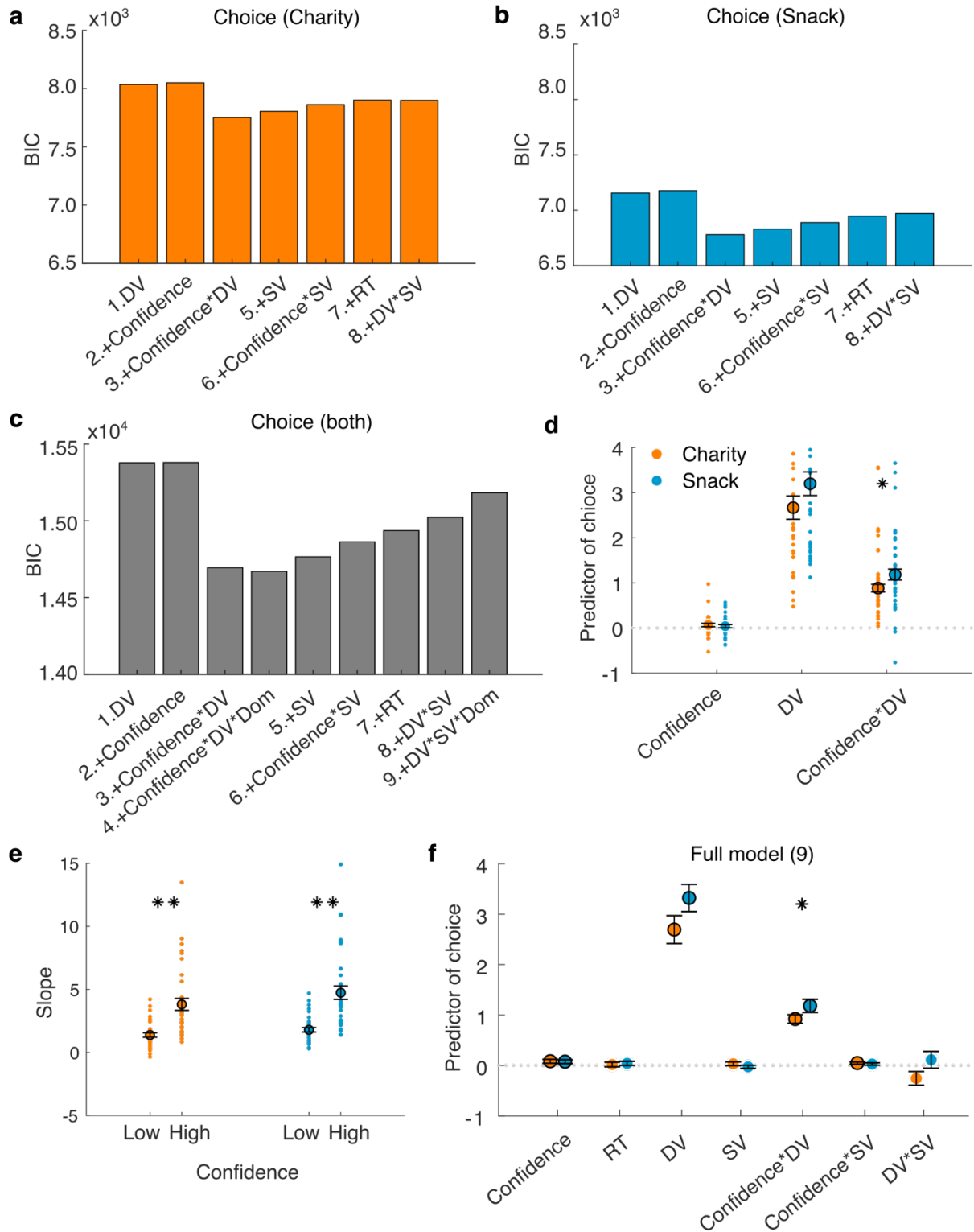
- in Cognitive Science*, 2(3), 511–527. <https://doi.org/10.1111/j.1756-8765.2010.01096.x>
- Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36(1), 163–177. <https://doi.org/10.1111/j.1551-6709.2011.01210.x>
- Pizarro, D. A., & Bloom, P. (2003). The Intelligence of the Moral Intuitions: Comment on Haidt (2001). *Psychological Review*, 110(1), 193–196. <https://doi.org/10.1037/0033-295X.110.1.193>
- Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3), 366–374. <https://doi.org/10.1038/nn.4240>
- Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications*, 10(1), 1–11. <https://doi.org/10.1038/s41467-019-09075-3>
- Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*, 1(May). <https://doi.org/10.1017/pen.2018.16>
- Schenk, T. (2006). An allocentric rather than perceptual deficit in patient D.F. *Nature Neuroscience*, 9(11), 1369–1370. <https://doi.org/10.1038/nn1784>
- Sepulveda, P., Usher, M., Davies, N., Benson, A., Ortoleva, P., & Martino, B. De. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *BioRxiv*, 2020.04.14.031971. <https://doi.org/10.1101/2020.04.14.031971>
- Shea, N., & Frith, C. D. (2019). The Global Workspace Needs Metacognition. *Trends in Cognitive Sciences*, 23(7), 560–571. <https://doi.org/10.1016/j.tics.2019.04.007>
- Shekhar, M., & Rahnev, D. (2020). Sources of Metacognitive Inefficiency. *Trends in Cognitive Sciences*, 1–12. <https://doi.org/10.1016/j.tics.2020.10.007>
- Skitka, L. J. (2010). The Psychology of Moral Conviction. *Social and Personality Psychology Compass*, 4(4), 267–281. <https://doi.org/10.1111/j.1751-9004.2010.00254.x>
- Soutschek, A., & Tobler, P. N. (2018). Motivation for the greater good: neural

- mechanisms of overcoming costs. *Current Opinion in Behavioral Sciences*, 22, 96–105. <https://doi.org/10.1016/j.cobeha.2018.01.025>
- Soutschek, A., & Tobler, P. N. (2020). Know your weaknesses: Sophisticated impulsiveness motivates voluntary self-restrictions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46, 1611–1623. <https://doi.org/https://doi.org/10.1037/xlm0000833>
- Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2005). Choosing the greater of two goods: Neural currencies for valuation and decision making. *Nature Reviews Neuroscience*, 6(5), 363–375. <https://doi.org/10.1038/nrn1666>
- Tarantola, T., Kumaran, D., Dayan, P., & De Martino, B. (2017). Prior preferences beneficially influence social and non-social learning. *Nature Communications*, 8(1). <https://doi.org/10.1038/s41467-017-00826-8>
- Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron*, 96(2), 348-354.e4. <https://doi.org/10.1016/j.neuron.2017.09.006>
- Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, 40(4), 1912–1920. <https://doi.org/10.1016/j.neuroimage.2008.01.057>

## 6. Supplementary figures



**Figure S1: Value domains' behavioural features.** **a-b.** Distribution of subjective values for each item rated in the first part of the experiment (Fig. 1a) on continuous scales from (-10) *really dislike* to obtain snack or support charity to (10) *really like*, respectively for charities and for snacks. **c.** Subjective parameters in both value domains being z scored within participants and across domains to observe differences between domain: |DV|= absolute difference in value between both items, SV = sum of values, Certainty = sum of certainties in values, RT = response time. **d.** Differences between both value domains in norms of performance: Coherence with one's subjective values and Consistency (repeated decision) over the two presentations of each pair of items. **e-f.** Response time as implicit uncertainty predicting both confidence (explicit uncertainty) and choice coherence for both value domains (paired t-tests\*: $p < 0.05$ , \*\*:  $p < 0.001$ ).



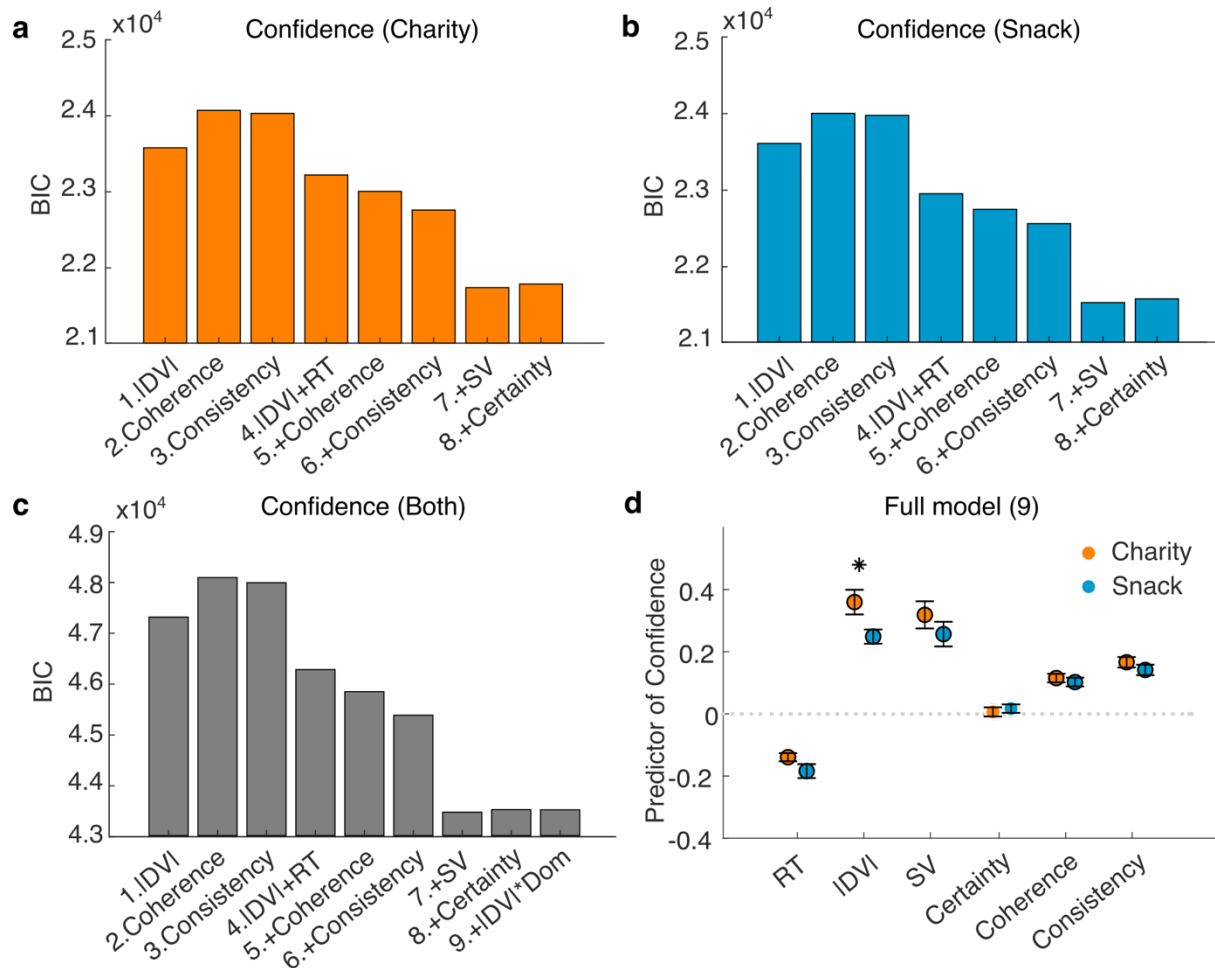
**Figure S2: Mixed models comparison for predictors of choice.** a-c. Mixed models were compared for their BIC either in each value domain or with both to test effect of domain interaction. Significant estimates ( $p < 0.05$ ) are circles in black and error

bar represents the SE of these estimates. The presence of a star represent the explanation of a difference between both domains by an interaction term in the corresponding full model tested for comparison (\*Dom: \*=p<0.05, \*\*=p<0.001). **d.** Best model of choice (c. number 4), where the interaction between Confidence and DV is also explained by the difference in domain. Interpretation must take into account the difference in difficulty and coherence between both domains (Fig. S1a,b,d). **e.** Modulation of choice sensitivity by confidence levels in both value domains: for each participant the sensitivity slope for low and high confidence (Fig. 1c-d) is significantly different (paired t-test: \*\*=p<0.001). **f.** Model 9 with all predictors tested highlights model 4 as best model due to other parameters' lack of significance in predicting the choice.

Full models of Choice tested in Sup Fig. 2c:

1. choseRside ~ 1 + zDV + (1 + zDV | subj)
2. choseRside ~ 1 + zDV + zConfidence + (1 + zDV + zConfidence | subj)
3. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV + (1 + zDV + zConfidence + zConfidence\*zDV | subj)
4. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom | subj)
5. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV | subj)
6. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV | subj)
7. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT | subj)
8. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV | subj)
9. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV\*zDom + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV\*zDom | subj)

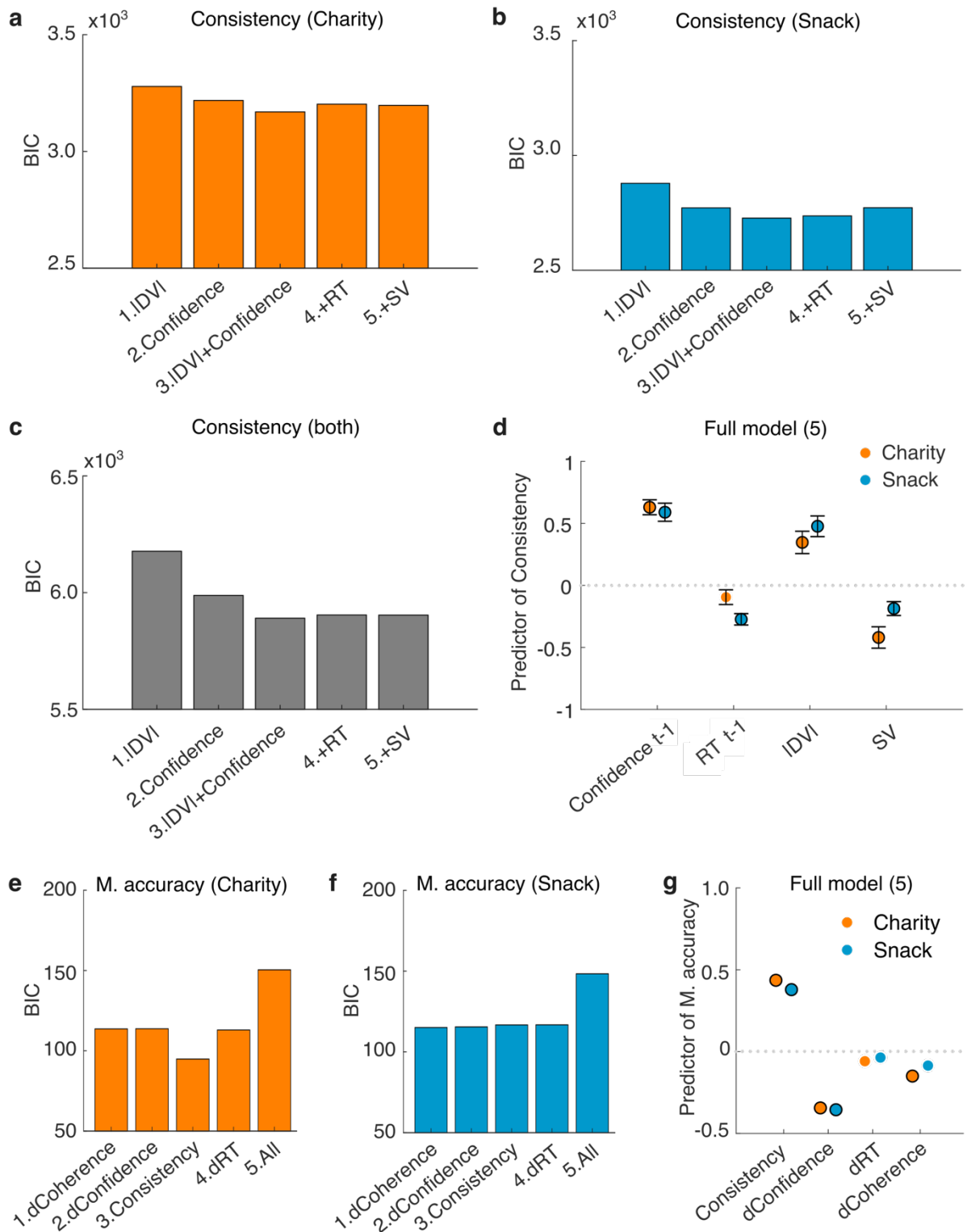




**Figure S3: Mixed models comparison for choice confidence.** a-c. Comparison of models predicting choice confidence for respectively Charity, Snack or Both (to test interaction of domain on predictor) values domains (model 6 is presented in Fig. 2a) d. Full and best model of confidence (model 8) with response time (RT), difference in value between both items(|DV|), sum of the items' value (SV), sum of certainty in the items' values (Certainty), choice of the high value item (Coherence), repeated choice (Consistency).

Full models of Confidence tested in Sup Fig. 3c:

1.  $z\text{Confidence} \sim 1 + z\text{aDV} + (1 + z\text{aDV} | \text{subj})$
2.  $z\text{Confidence} \sim 1 + z\text{Coherence} + (1 + z\text{Coherence} | \text{subj})$
3.  $z\text{Confidence} \sim 1 + z\text{Consistency} + (1 + z\text{Consistency} | \text{subj})$
4.  $z\text{Confidence} \sim 1 + z\text{aDV} + z\text{RT} + (1 + z\text{aDV} + z\text{RT} | \text{subj})$
5.  $z\text{Confidence} \sim 1 + z\text{aDV} + z\text{RT} + z\text{Coherence} + (1 + z\text{aDV} + z\text{RT} + z\text{Coherence} | \text{subj})$
6.  $z\text{Confidence} \sim 1 + z\text{aDV} + z\text{RT} + z\text{Coherence} + z\text{Consistency} + (1 + z\text{aDV} + z\text{RT} + z\text{Coherence} + z\text{Consistency} | \text{subj})$
7.  $z\text{Confidence} \sim 1 + z\text{aDV} + z\text{RT} + z\text{SV} + z\text{Coherence} + z\text{Consistency} + (1 + z\text{aDV} + z\text{RT} + z\text{SV} + z\text{Coherence} + z\text{Consistency} | \text{subj})$
8.  $z\text{Confidence} \sim 1 + z\text{aDV} + z\text{RT} + z\text{SV} + z\text{Scert} + z\text{Coherence} + z\text{Consistency} + (1 + z\text{aDV} + z\text{RT} + z\text{SV} + z\text{Scert} + z\text{Coherence} + z\text{Consistency} | \text{subj})$
9.  $z\text{Confidence} \sim 1 + z\text{aDV} * z\text{Dom} + z\text{RT} + z\text{SV} + z\text{Scert} + z\text{Coherence} + z\text{Consistency} + (1 + z\text{aDV} * z\text{Dom} + z\text{RT} + z\text{SV} + z\text{Scert} + z\text{Coherence} + z\text{Consistency} | \text{subj})$



**Figure S4: Mixed model comparison for choice consistency and subjective metacognitive accuracy.** **a-c.** Mixed models comparison for choice consistency in second presentation of the pair of items with confidence at the previous encounter of the same pair of items, respectively for each value domains and both to test for domain interaction. Confidence and response time are taken at t-1 whereas

difference of value and sum of value are common to the item pairs both at t-1 and t, model 4 is presented in Fig. 2b. **c.** Full model (5) of consistency from which  $\beta$  Confidence t-1 is taken to calculate the correlations in Fig. 3c (as in Folke et al., 2017). **e-f.** Mixed model comparison for participants' metacognitive abilities as measured from best model in Fig. S2d comparing individual behavioural parameters as candidate predictor of this subjective insight, and all parameters combined in model 5 (see models below). **g.** Fixed effects of model 5 as predictors of metacognitive accuracy in both domains for behavioural differences between first and second presentation of each choices to study behavioural optimisation over time: difference in average Coherence between first and second presentations (dCoherence), difference in average confidence (dConfidence), score of repeated choices (Consistency), difference in average response time (dRT),

Full models of Choice consistency over the 2 presentations tested in Fig. S4a-c:

1. Consistency ~ 1 + zaDV + (1 + zaDV | subj)
2. Consistency ~ 1 + zConfidence + (1 + zConfidence | subj)
3. Consistency ~ 1 + zaDV + zConfidence + (1 + zaDV + zConfidence | subj)
4. Consistency ~ 1 + zaDV + zConfidence + zRT + (1 + zaDV + zConfidence + zRT | subj)
5. Consistency ~ 1 + zaDV + zConfidence + zRT + zSV + (1 + zaDV + zConfidence + zRT + zSV | subj)

Full models of subjective metacognitive accuracy tested in Fig. S4e,f:

1. metaAcc ~ 1 + dCoherence + (1 + dCoherence | subj)
2. metaAcc ~ 1 + dConfidence + (1 + dConfidence | subj)
3. metaAcc ~ 1 + Consistency + (1 + Consistency | subj)
4. metaAcc ~ 1 + dRT + (1 + dRT | subj)
5. metaAcc ~ 1 + dCoherence + dConfidence + Consistency + dRT + (1 + dCoherence + dConfidence + Consistency + dRT | subj)

















## 7. Supplementary material

### List of stimuli:

#### Charities

 <p>Acting for hunger relief</p>	 <p>Providing education</p>	 <p>Preventing climate change</p>	 <p>Refuting climate change claims and regulations</p>
 <p>Preventing LGBT rights</p>	 <p>Acting for LGBT rights</p>	 <p>Acting for species conservation</p>	 <p>Advancing cancer research</p>
 <p>Advancing genetically modified food</p>	 <p>Preventing genetically modified food</p>	 <p>Providing disaster medical care</p>	 <p>Providing housing for all</p>
 <p>Acting for women's rights</p>	 <p>Acting for culture and art preservation</p>	 <p>Supporting euthanasia (assisted suicide)</p>	 <p>Preventing euthanasia (assisted suicide)</p>

Snacks

 <p>Caramel and nougat chocolate bar</p>	 <p>Roasted beans</p>	 <p>Smoked and cooked beef</p>	 <p>Liquorice candies assortment</p>
 <p>Surprising flavors jelly beans</p>	 <p>Dried mango</p>	 <p>Fruit flavored candies</p>	 <p>Chocolate and orange protein bar</p>
 <p>Crunchy and salted pork rinds</p>	 <p>Salted potato chips</p>	 <p>Salted potato chips</p>	 <p>Caramel flavored rice cakes</p>
 <p>Orange flavored dark chocolate</p>	 <p>Wheat and banana bar</p>	 <p>Dried fruits and nuts mix</p>	 <p>Mixed flavored candies</p>



# Bridge:

## From moral to general.

This thesis argues for the role of subjective value in providing a comprehensive picture of procedural metacognition as a thermostat for decision coherence. By developing the models of both the function and the computation of metacognitive monitoring signals, we suggest that subjective value is essential to close the loop and provide a comprehensive understanding of metacognition. The second half of the thesis concerns empirical work that aims at providing new models for the computation of confidence reports in regard to subjective value.

Both Chapters 4 and 5 were concerned with the contribution of value input into a decision's confidence level, whether it be based on limited knowledge and heuristic cues or on moral values that one might be very explicitly aware of as part of her subjective identity. If both chapters suggest that metacognition acts as a coherence thermostat by using subjective value both to guide a decision and to reflect upon its reliability retrospectively, therefore covering a wide array of subjective values domains, one can ask: if value enables ubiquitously to inform confidence reports as to the decision coherence, is this metacognitive thermostat general to the individual and ubiquitous to all her decisions? Would there be a subjective trait or metacognitive signature as a criterion for this thermostat to be set at a certain cognitive and behavioural "temperature" of coherence?

Chapter 6 starts by testing the potential of confidence signals to track, as a thermostat would, not whether the decision is correct or wrong but in a more continuous way whether it be sensitive to the value of the item itself.

Chapter 6 offers to step away from the question of what metacognition tracks in a choice to how it tracks overall. If metacognition indeed acts as a thermostat or a coherence monitor, then one would expect it to be set with a certain sensitivity or criterion for this coherence. Across two value domains, Chapter 6, therefore, asks

whether metacognitive monitoring appears rather optimal (i.e. tracking the best option amongst the available ones) or rather satisfiable (i.e. tracking the value of the chosen item itself). While defining this metacognitive style within an individual, as a criterion of sensitivity to coherence, one can then ask: is this metacognitive style general across value domains? In other words, is there something subjective and ubiquitous to the agent that defines her sensitivity to decision coherence?



## Chapter 6

# Beyond the rational monitoring of value-based decisions.

### Abstract

Are you sure you prefer to support biodiversity than environmentalism? And you that you want the dark chocolate over the milk chocolate ice cream? Here we ask what does confidence levels in value-based choices rely on. Indeed, in choices where there is an objective accuracy to follow in our choices, it can be argued that confidence should reflect the probability of this latter to be correct, in a normative manner. But in preferential choices that rely on continuous and subjective values, is confidence also optimal by tracking how much better the chosen option is over the alternative? Or does one only evaluate her choice based on the value of the chosen item in a satisfiable manner? To answer this question, we asked participants to rate their likings of individual snacks and charity before asking them to choose amongst pairs of one or the other and report their confidence in having chosen their favourite item. With our small sample of N=36 participants, we obtained preliminary results on subjective metacognitive profile across value-domains. First, we observed a negative correlation in both moral and hedonic domains between the participants' metacognitive optimality and satisfiability: in other words, we observe distinct profile whereby some participants appeared to rather track how optimal their choices were whereas for some others, confidence instead seemed to focus on tracking the satisfiability of the chosen item. These preliminary results suggest a distinctive metacognitive profile as being rather metacognitively optimal or

satisfiable. Lastly we observed that across both value domains, subjective optimality or satisfiability appeared to be maintained as presenting a general subjective trait across value domains. While these results are preliminary, we discuss how value-based domain can offer new predictors of confidence to understand individuals through their subjective metacognitive fingerprint.

*Keywords:* confidence, value-based choices, preference, metacognition, moral, hedonic.

## **1. Introduction**

Do I prefer to support an environmental or a social cause? Am I sure this is the right thing to do? As humans, our ability to reflect on our own views and actions defines us as rational moral agents, because it supports either moral reasoning and regret, or self-evaluation in general (Paxton & Greene, 2010). This capacity to reflect upon our own choices has been operationalised in other domains as a metacognitive ability which assesses whether our decisions are coherent with a given norm, for instance objective instructions (Pouget et al., 2016). Confidence ratings give us an insight in how individuals reflect and subjectively evaluate their own choices given this norm. In economic choices (for example, a two alternative forced choice between snacks (Folke et al., 2017) or everyday objects (De Martino et al., 2017)), participants report higher confidence when they choose the item they like best, and lower confidence when they find themselves choosing the less liked item. In other words, if an individual said that she really liked a chocolate bar and just tolerated crisps, she would be confident in her choice if she chooses the chocolate bar over the crisps when presented with both options (De Martino et al., 2013). On the contrary, if she chose the crisps, she would report low confidence in having chosen her favourite item. Therefore, even in choices where norms are indexed on one's subjective preferences rather than on an objective rule, participants can reflect on whether their choices are consistent with a norm, in a discriminate manner. This reflective ability also relates to the ability to adjust one's behaviour over time by making choices which are more coherent with one's values (Folke et al., 2017).

We previously extended this finding to a new domain of value-based decisions: moral choices. Indeed, we found that participants were able to reliably report with their feeling of confidence whether they chose the charity to which they attributed the highest subjective value, and do so with an overall comparable discriminability as in hedonic choices.

In the present experiment, we explored whether confidence in moral choices could, as confidence in hedonic choices, be satisfied by valuable but suboptimal choices. Indeed, an interesting finding in economic tasks is that confidence does not only track whether a choice maximizes one's preferences but also whether the choice is likely to be of high value to the individual, and so even if it is sub-optimal (Hertz et al., 2018). We therefore investigated the metacognitive monitoring rule when individual evaluated their choices in regard to their moral preferences: either tracking choice optimality or satisfiability.

To test these hypotheses, we first asked participants to rate how much they valued both a set of snacks and a set of charities. They then saw pairs of snacks or pairs of charities and had to choose the item they preferred, before reporting how confident they were that they had chosen their favourite option. Building on our previous findings that humans have insight into the coherence of their hedonic and moral choices, we investigated whether in a continuous manner, confidence was also cued on the magnitude of subjective values. More precisely, we tested whether confidence would track, beyond the optimality of the choice, whether the choice was likely in average to satisfy one's standards. Our results highlight that in the moral as in the hedonic domains, confidence is, further than tracking whether the decision was optimal by choosing the most value option, it also tracks the value of the chosen item for its own sake and independently from the alternative items.

## **2. Method**

### ***2.1. Participants***

A total of 43 participants fluent in English were recruited to take part in the computer-based experiment at the LMU's psychology laboratory. To test metacognitive ability in Charity choices, we used a similar sample size to studies

demonstrating metacognitive abilities in Snack choices (Folke et al., 2017). All participants signed a consent form and were compensated 9 euros per hour with a 1 euro bonus if they performed above 85% in the attention checks and comprehension questions at the end of the experiment. We rejected 7 participants whose choices' coherence score (according to reported likings) was out of range 60-95%, in either value domain. The remaining 36 participants were included in the analyses (17 females, aged: 20-43). The study was approved by the University of London Research Ethics Committee (Project Number: SASREC\_1819-313A).

## **2.2. Stimuli**

A total of 16 snacks were selected for their differences and similarities along 4 axes: sweet-savoury, healthy-rich, single item- pack of items and rare-popular snack. For charities, the selection was based on Maoz et al., 2019, from which 8 consensual charities and four pairs of controversial charities were selected. This selection was made to allow similarities and differences in six categories of causes: conservation (of environment and biodiversity), (human) rights, health (and research), support (to people in need), (social) inequalities and culture (and education). High definition pictures of the snacks and of the names and logos of the charities were downloaded from the internet. Finally, in order to normalize the information available for all stimuli, we applied a standard format of 400x400 pixels and added a few words of description of the snack or of the charity's aims.

## **2.3. Procedure**

After reading general instructions and consent forms, participants saw a snapshot of all the options used in the experiment (8 at a time for 20 seconds) and were asked to familiarise themselves with the range of options.

*Liking.* After a 500ms fixation cross, participants saw either a snack or a charity and rated how much they would like to respectively *obtain* this snack or *support* this charity on a continuous scale from *really dislike* to *really like* (quantified as -10 to 10 step .5 Fig. S1a-b, Lebreton et al., 2015). They then rated how certain they were of this rating, on a continuous scale from *not at all* to *absolutely certain* (Fig. 1a).

Participants saw in a random order 2 blocks of either 16 snacks or 16 charities, which they also came across in a random order.

*Choice.* In each trial, participants saw a 500ms fixation cross, and then either two snacks or two charities. They were asked to select the item they would prefer to respectively *obtain* or *support*. After each choice, participants were asked how confident they were that they chose their preferred item from *not at all* to *absolutely* confident (quantified for analysis as 0 to 10 step .1). No time pressure was applied in the task. All combinations of two snacks or two charities (N=120 each) were presented twice in reversed lateral position (480 trials). The full task was designed in JavaScript (JSpsych.org) and consisted of 8 blocks in a random order presenting a series of 60 random choices of a same value domain. A five minutes demo of the task is available at:

<https://www.cvbe-experiments.com/oa/MoralConfidenceDemo/>

### 3. Results

While decisions are often evaluated in binary manner for being accurate in perceptual choices or coherent in value-based decision, here we compare two continuous predictors. In behavioural economics, a common concept is choice satisfaction whereby individuals do not thrive for coherence if they can select another item which they appreciate enough to their standards (high value item). In a monetary task, confidence was also demonstrated to partly track such incoherent decision when the chosen item was of high value (Folke et al., 2017; Hertz et al., 2018). We therefore ask here whether confidence in moral choices would also follow incoherent choices when both options are of high value to the participant.

#### 3.1. Items value and choice coherence.

*Satisfying moral choices.*

First, we observed decision making behaviour in regard to the norm of coherence: choosing the most valued item. We found that |DV| was the best predictor of coherent moral choice (mixed model Fig. 1a Charity:  $\beta=0.74$ ,  $se=0.11$ ,  $z=6.48$ ,  $p<10^{-10}$ , Snack:  $\beta=0.88$ ,  $se=0.08$ ,  $z=10.41$ ,  $p<10^{-24}$ ). Secondly, the same model found that the average value of the charities to choose from (SV) was a significant negative factor

for choice coherence (Charity:  $\beta=-0.28$ ,  $se=0.10$ ,  $z=-2.68$ ,  $p<0.01$ ). In other words, when both charities were on average highly valued, participants were likely to be satisfied by any of both charities and not to thrive to select the charity which they rated with higher value. Furthermore, we found a negative interaction between both DV and SV, known as the Weber and Fechner law (Charity:  $\beta=-0.25$ ,  $se=0.12$ ,  $z=-2.10$ ,  $p<0.05$ ). This finding suggests that the more valuable charities were in average, the more difference in value was required for participants to choose their favourite charity.

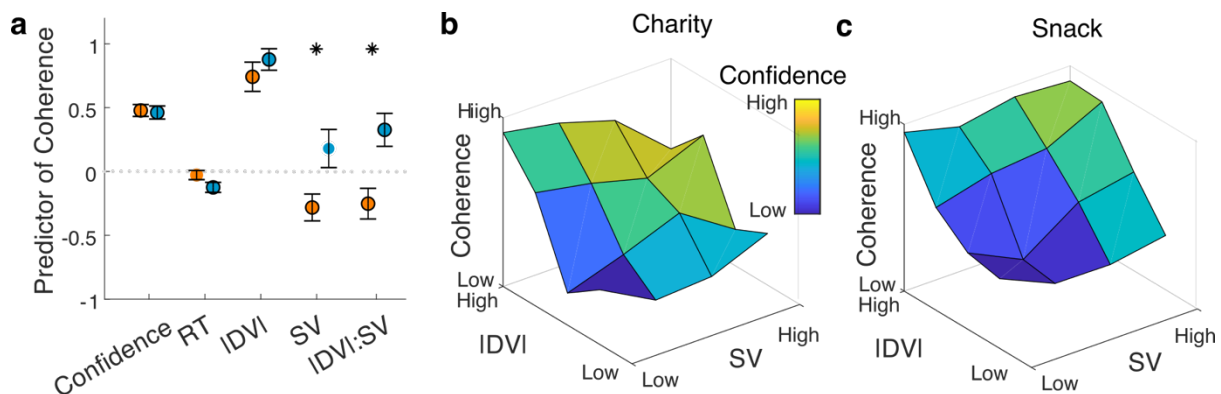
### *Unsatisfying hedonic choices.*

We also investigated the effect of items values on choice coherence for the hedonic domain. Interestingly, in hedonic choices we did not find an effect of SV on choice coherence ( $\beta=0.18$ ,  $se=0.15$ ,  $z=1.20$ ,  $p=0.23$ ) and against the predictions from previous literature (Folke et al., 2017), we found the opposite positive interaction of DV and SV in snack choices ( $\beta=0.33$ ,  $se=0.13$ ,  $z=2.53$ ,  $p<0.05$ ). Therefore in our experiment, the more valuable the snacks were in average, the more participants appeared to thrive to choose their favourite item. This difference between both domains in choice satisfaction was highlighted by our mixed model (significant interaction in Fig 1a, domain x SV:  $\beta=0.23$ ,  $se=0.10$ ,  $z=2.24$ ,  $p<0.05$ , domain x SV x DV:  $\beta=0.28$ ,  $se=0.09$ ,  $z=2.95$ ,  $p<0.01$ ). From our results we can therefore conclude for the least that when both options are appealing, participants appear to have a decreased sensitivity to choose their favourite charity compared to their favourite snack, hence highlighting a satisfying behaviour in moral choices (Fig. 1b: coherence decreases with increasing SV).

### **3.2. Satisfied confidence.**

Finally, we investigated the relation of confidence to these incoherent and satisfying choices. In the experiment, participants were repeatedly asked at each block to report their confidence in “having chosen their favourite item” and successfully reported this instruction at the end quiz. With this design, we aimed at testing participants’ ability to reflect upon the coherence of their choice, namely testing their insight in the binary evaluation of their decision as having (or not) chosen their

most valued item. Accordingly, we expected that even if a choice would be satisfying participants' values, participants would still report any incoherent choice with low confidence, thereby demonstrating an unbiased metacognitive insight. In both Charity and Snack choices however, we found that confidence was sensitive not only to |DV| (predicting coherent choices) but also to SV (Fig S3d: Charities |DV|:  $\beta=0.36$ ,  $se=0.04$ ,  $z=9.06$ ,  $p<10^{-18}$ , Snack |DV|:  $\beta=0.25$ ,  $se=0.02$ ,  $z=2.27$ ,  $p<10^{-26}$ , Charities SV:  $\beta=0.31$ ,  $se=0.04$ ,  $z=7.29$ ,  $p<10^{-15}$  Snack SV:  $\beta=0.26$ ,  $se=0.04$ ,  $z=6.47$ ,  $p<10^{-12}$ , similar effect of SV in both domains: domain x SV:  $\beta=0.00$ ,  $se=0.01$ ,  $z=0.46$ ,  $p=0.64$ ). We therefore observed in Charity choices that confidence was therefore increasing with SV while choices coherence was dropping (Fig 1c). Although we could not test for the actual effect of satisfying choices in the snack choices, we nonetheless found a metacognitive blindsight in Charity choices where participants reported by their high confidence the belief of having chosen coherently when they did not. These results therefore highlighted that in value based choices metacognition appears not only to be tuned to track choice coherence but also the likelihood of a choice to be of high value, and so regardless of the alternative item. Such value-based norm of confidence was therefore tested in the following part.



**Figure 1: Choice coherence and confidence.** **a.** Predictors of coherent choices (model 7 from mixed models comparison Fig. S3d), namely confidence in the choice, response time, difference in items value, overall items' value and interaction in the latter two. For both value domains, significant predictors are contoured in black, error bar represents the SE of these estimates, and significant interactions between both domains are represented by \* when significant ( $p<0.05$ ). **b-c.**

Summary of the interaction between the four main variables: choice coherence as predicted by DV and SV and their relation to choice confidence.

### ***3.3. Value-based norms for confidence.***

*Testing norms.*

Lastly we investigated which parameters of the decision predicted confidence. Mainly, we compared two orthogonal norms which confidence could track about a decision:

$$\text{choice optimality} = \text{value}_{\text{chosen}} - \text{value}_{\text{unchosen}}$$

$$\text{choice satisfiability} = \text{value}_{\text{chosen}}$$

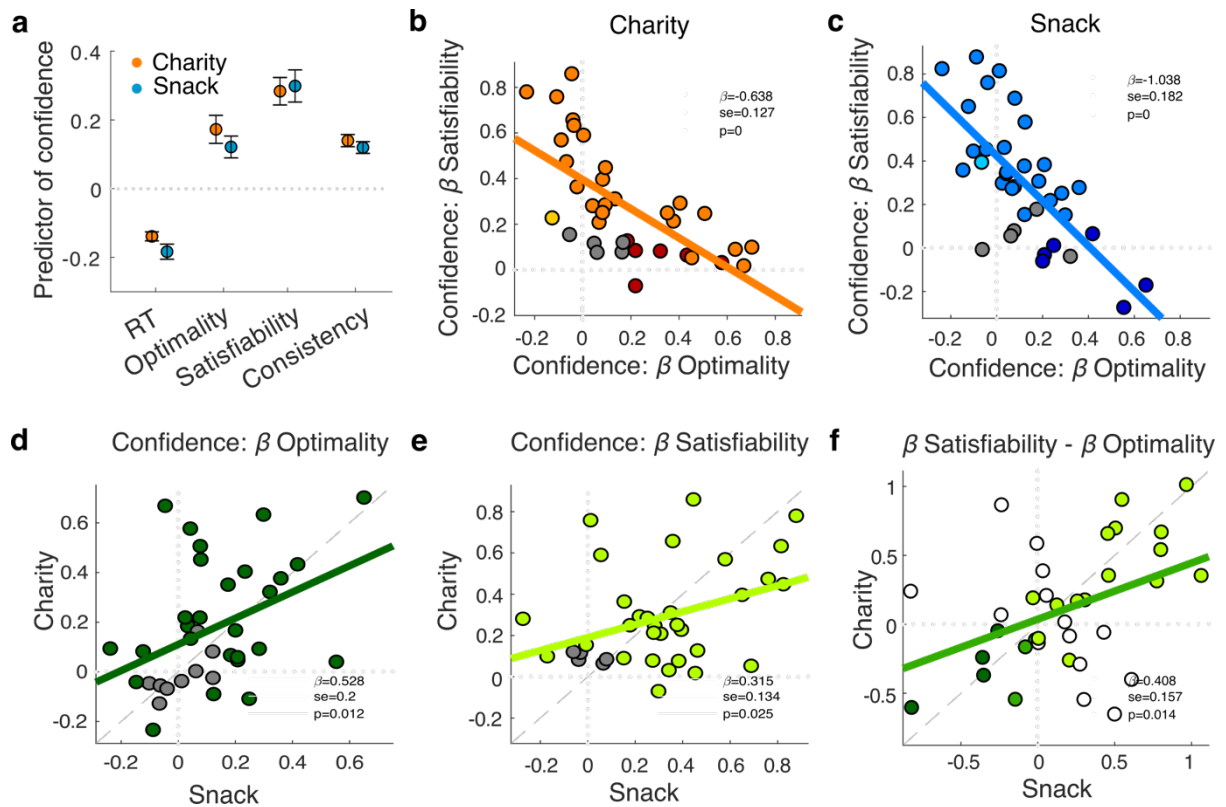
As for the norm of coherence, the norm of optimality (as defined in this paper) is relative a relative one as it takes into account the value of both items available in the choice. The difference between both norms lies in the continuous nature of our norm of optimality, being therefore more in line with the nature of the value-based task. With this norm of optimality, we suggest that confidence tracks, further than whether a choice did follow (or not) the individual's preferences (norm of coherence), confidence also tracks by how much the chosen item was better than the alternative. On the contrary, our norm of choice satisfiability detaches from the relative aspect (that confidence in perceptual tasks takes with the norm of accuracy), making our norms of satisfiability and of optimality orthogonally independent. Instead, inspired from value-based choices literature (Teodorescu et al., 2016), our norm of choice satisfiability is absolute by only considering the chosen item. With this norm, we suggest that retrospective confidence therefore only tracks how valuable the chosen option is, and does so regardless to the alternative item. Hereby propose that, as previously argued for the decision process, value-based metacognition might too not be tuned to be optimal, but rather be satisfiable by disregarding some of environment's information when the choice promises to satisfy the individual's values. Consequently, independently from the decision process which would lead to a sub-optimal satisfying choice, our metacognitive norm of choice satisfiability proposes that the individual reflecting on such choice would not consider the choice as sub-optimal, but instead would only evaluate how good the chosen item is for its own sake.



With a mixed model, we investigated which of four independent norms of choice were best monitored by confidence (Fig. 2a). Our results show that both choice optimality and satisfiability contributed to confidence (Charity optimality:  $\beta=0.17$ ,  $se=0.04$ ,  $z=4.23$ ,  $p<10^{-4}$ , Snack optimality:  $\beta=0.12$ ,  $se=0.03$ ,  $z=3.81$ ,  $p<10^{-3}$ , Charity satisfiability:  $\beta=0.28$ ,  $se=0.04$ ,  $z=7.10$ ,  $p<10^{-11}$ , Snack satisfiability:  $\beta=0.30$ ,  $se=0.05$ ,  $z=6.39$ ,  $p<10^{-9}$ ). Therefore, against the common concept that confidence tracks the relative validity of a decision, here our results suggest that in both our value domains, value-based metacognition also monitors how satisfiable a choice is for itself. These results align with our previous findings about confidence in satisfied choices (part 3.2) and suggests that confidence could also be high in satisfied and sub-optimal (incoherent) choices in the hedonic domain.

#### *Norms relation.*

To refine our understanding of how these norms contribute to the metacognitive function, we investigated at the individual level the relation between confidence's sensitivities to choice optimality and satisfiability. In other words, through this post hoc analysis, we aimed at testing whether, for value-based tasks, the metacognitive process was equally built on both norms or whether some individuals' metacognition could tend toward tracking either choice optimality or choice satisfiability. In both value domains, we found despite our small population size strong negative relations between confidence's sensitivities to choice optimality and confidence's sensitivity to choice satisfiability (Fig 2b,c, linear regressions Charity DV:  $\beta=-0.64$ ,  $se=0.13$ ,  $z=-5.01$ ,  $p<10^{-4}$ , Snack DV:  $\beta=-1.04$ ,  $se=0.18$ ,  $z=-5.71$ ,  $p<10^{-5}$ ). We can notice that, while most participants' confidence tracked both of the norms (middle orange or blue) some participants appeared to have either optimal (dark orange or blue) or satisfiable (bright orange or blue) metacognition. Namely, in our value-based choices tasks, some individuals' reflective process tended to be more absolute and focused on the value of their decision and while other tended to be more optimal by always keeping in mind the other option which was available to them.



**Figure 2: Norms of confidence in value based choices.** **a.** Norms predicting confidence, namely response time, choice optimality (the difference in value between chosen and unchosen item), choice satisfiability (value of the chosen item) and consistency (repeated decisions among both presentation of items pair). Best model of predictors for choice confidence (model 7 from mixed models comparison Fig. S3c) presenting for both value domains significant predictors as contoured in black and error bar for SE of these estimates. **b-c.** Linear regression between the subjective sensitivities of confidence to choice optimality and choice satisfiability in both value domains, as taken from model in panel a. **d-e.** Linear regression testing the stability across value domains of confidence sensitivities to the norms of either (d) optimality or (e) satisfiability within individuals. **f.** Linear regression demonstrating stable metacognitive profile as either being tuned to choice optimality or satisfiability across value domains. Colour coding represents the metacognitive profile across value domains as participants being metacognitively stable in both value domains at having a significant sensitivity to either choice optimality (dark green), satisfiability (bright green), both (middle green), or none (white).

### 3.4. Domain generality

Since our measures of value-based metacognition appeared to capture a subjective trait of metacognition, we then pushed this finding further by testing whether confidence optimality and satisfiability would be further than domain specific, a subjective trait shared across value domains. In other words, following the domain generality hypothesis of metacognition (Rouault et al., 2018), we investigated whether metacognitive sensitivities in different tasks appeared to belong to one same monitoring system, or instead to different metacognitive systems tuned for their own value domains. First of all, we tested whether the traditional measure of confidence's sensitivity to the norm of coherence (part 3.1, Fig. S2d) presented such a domain generality trait across domains and did not find any significant link with our sample size (N=36, Fig. S3d:  $\beta=0.23$ ,  $se=0.15$ ,  $z=1.50$ ,  $p=0.14$ ). We then looked at the sensitivities of confidence to track both choice optimality and choice satisfiability across our two tasks and found in both cases a significant link of these abilities within participants (Fig 2d,e, Optimality:  $\beta=0.53$ ,  $se=0.20$ ,  $z=2.64$ ,  $p<0.05$ , Satisfiability:  $\beta=0.31$ ,  $se=0.13$ ,  $z=2.34$ ,  $p<0.05$ ). Hence, participants' abilities to report through their confidence choice optimality and choice satisfiability in one value domain, was likely to predict these confidence sensitivities in the other value domain. To push these findings further, we then wondered whether a metacognitive profile of satisfiability would be conserved across value domains. Specifically, while participants seemed to be either rather metacognitively optimal or satisfiable (Fig 2b,c) and that both seemed to be subjective trait across domains (Fig. 2d,e), we then tested whether participants who were rather sensitive to choice satisfiability (compared to optimality) in one domain also tended to be so in the other domain. Both our measures of metacognitive sensitives being generated by the same model which accounted for each other made them comparable measures, hence allowing us to create for each participants a metacognitive profile based on a subtraction between both measures. Although some participants' metacognition was not significantly stable at tracking either norms across domains (white), we found a significant general link among our participants by being either metacognitively stable at tracking choice optimality, satisfiability, or both across domains ( $\beta=0.41$ ,  $se=0.16$ ,  $z=2.60$ ,  $p<0.05$ ). Together our results shed light on the flexibility of the

metacognitive process when reflecting upon value-based choices both within and across individuals. While previous studies of metacognition defined our reflective process as monitoring a simplified binary relative norm (such as accuracy or coherence), here we define a more realistic and pragmatic metacognitive profile based on continuous norms.

#### **4. Discussion**

Following prospect theory of decision making in economics (Odhoff, 1965), we explored whether at the metacognitive level, participants could be calibrated to follow a non-optimal norm as by having a confidence tracking choices satisfied with high value items. Indeed, our results define for the first time to our knowledge, an absolute norm for confidence in choice: here we discuss the implications of having a metacognitive sensitivity to choice satisfiability.

Although we explicitly asked participants to report choice coherence through their confidence, it could be debated whether participants tracking choice satisfiability did so by inability or as a voluntary strategy. On the one hand, individuals could consciously decide to only strive for decisions satisfying a standard which they consider as optimal enough and by the same conscious strategy also disregard this satisfied sub-optimality in their reflection process. On the other hand, metacognitive satisfiability could be a true metacognitive blindsight by which they could not evaluate the sub-optimality of their choice when it is of high value. Evaluating subjective traits as perfectionism and self-knowledge (to define one's standards) might bring insight in whether this sensitivity to choice optimality is a metacognitive strategy or blindsight.

At the computational level, the race model of decision making was previously used to gain insight in the cerebral buildings of confidence levels (De Martino et al., 2013). Nonetheless, such relative model based on the consideration of both items to build a confidence level does not take into account the absolute norm of choice satisfiability. Future computational studies could therefore apply refined models to define a more complete picture of the buildings of confidence levels. For instance, an investigation of the shared metacognitive sensitivity to the absolute norm of choice satisfiability in both a value-based and a perceptual task could use models

---

from perception while investigate further the domain generality hypothesis suggested by our small sample size. (Bang & Fleming, 2018; Hertz et al., 2018).

Furthermore, the link between metacognitive sensitivity to satisfiability and the decision rule could be investigated in future studies. Indeed, while the behavioural economics literature defines a bounded rationality at the decision level, to our knowledge, no studies have been made to study the role of metacognition in such behavioural trait. While bounded rationality describes how individuals will choose sub-optimally as to save cognitive resources, time or effort, the consideration of a bounded metacognition could shed light in the nature of the decision rule. Indeed, if the individual has no ability to reflect upon his choice optimality, an individual might be limited to choosing according to a satisfying rule, while if the participant have the reflective ability to identify choice sub-optimality, she might rather decide to choose to satisfy in given contexts as a choice strategy while being aware of a possibility to optimize choices in other tasks and contexts.

At the subjective trait, we also found that participants generally tracked partly both choice optimality and choice satisfiability with their confidence and often tended to track one norm more than the other in both value domains. Although our study was performed on a limited population size (N=36), such subjective differences in the norm which metacognition is tuned to monitor raises questions about the implications of such variable sensitivities. For instance on the one hand, having a metacognitive satisfying profile could be seen as a limitation to optimize behaviour. Indeed, confidence is defined as a guiding norm for learning in absence of feedback and the question of whether participants with a satisfiable metacognitive profile could also have sub-optimal learning ability in environments with high value items arises. Namely, if participants tuned to detect choice optimality can identify the opportunity to improve their decision in the future, but that participants tracking choice satisfiability are not aware of such opportunity, does that imply that their ability to learn and optimize their behaviour would be more limited? On the other hand, could participant who's metacognitive ability is tuned only to choice optimality always thrive to optimize their choices until they reach the highest item? In a society globalized and capitalistic society where options can be overwhelming,

could individual with optimal metacognitive profile be more subject to anxiety, burn out and depression? As discussed before, if subjective traits as self-knowledge could contribute to the elaboration of metacognitive sensitivity to choice satisfiability, could that modify the way optimizers would evaluate poorly high value choices because of their sub-optimality?

The societal implications of metacognitive ability in well-being and psychological disorders is a growing interest. Nonetheless, to this day this research has mainly assesses the relation of metacognition to cognitive disorders through psychometric tasks which are far distant from the struggle that face this population. Here, our findings support a domain general metacognitive ability with a profile between reflective sensitivity to choice optimality to choice satisfiability. Through this seemingly shared metacognitive resource across value-based task, our results support the existing consideration that, if metacognition is not the cause of behavioural and evaluative psychological disorders, it could for the least contribute to help this population through therapies (Bhome et al., 2019; Faivre et al., 2019; Heyes et al., 2020; Lysaker et al., 2014; Vaghi et al., 2017). Furthermore, various psychological theories suggest altruistic and moral behaviour to be necessary to one's well-being (Frankl, 1959; Seligman & Csikszentmihalyi, 2000). Our findings that humans withhold the conscious knowledge about being or not behaviourally aligned with their moral views therefore also supports these applications of moral pursuits to the non-clinical population. Through this experiment, we aimed at extending the application of the study of metacognition to a more ecological context, hereby hedonic and moral choices, to support the extension of this promising field to a more ecological settings and its possible societal contributions.

## **5. Conclusion**

In the present study, we demonstrated that both in moral and hedonic choices, participants track, further than whether a choice is optimal, also whether it is up to their standards. Altogether our results paint a more comprehensive picture of how metacognition operates in ecological value based choices by tracking both choice optimality and satisfiability while serving as a guide for future behaviour.

### **Author contributions**

O.A. conceptualised and designed the study, collected and analysed the data, wrote the manuscript, B.D.M. supervised and guided data analysis, O.D. supervised and funded the data collection. All authors interpreted the data and revised the manuscript.

### **Acknowledgments**

We wish to thank the Graduate School for Systemic Neurosciences LMU for supporting this doctoral project, Lucas Battich for his technical guidance in designing the study and Justin Sulik for his insight in the data analysis.

### **Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

### **Data availability**

The data presented in this article can be found on OSF at:

[https://osf.io/xatrz/?view\\_only=810578017c2c43c19342aab589b70274](https://osf.io/xatrz/?view_only=810578017c2c43c19342aab589b70274)

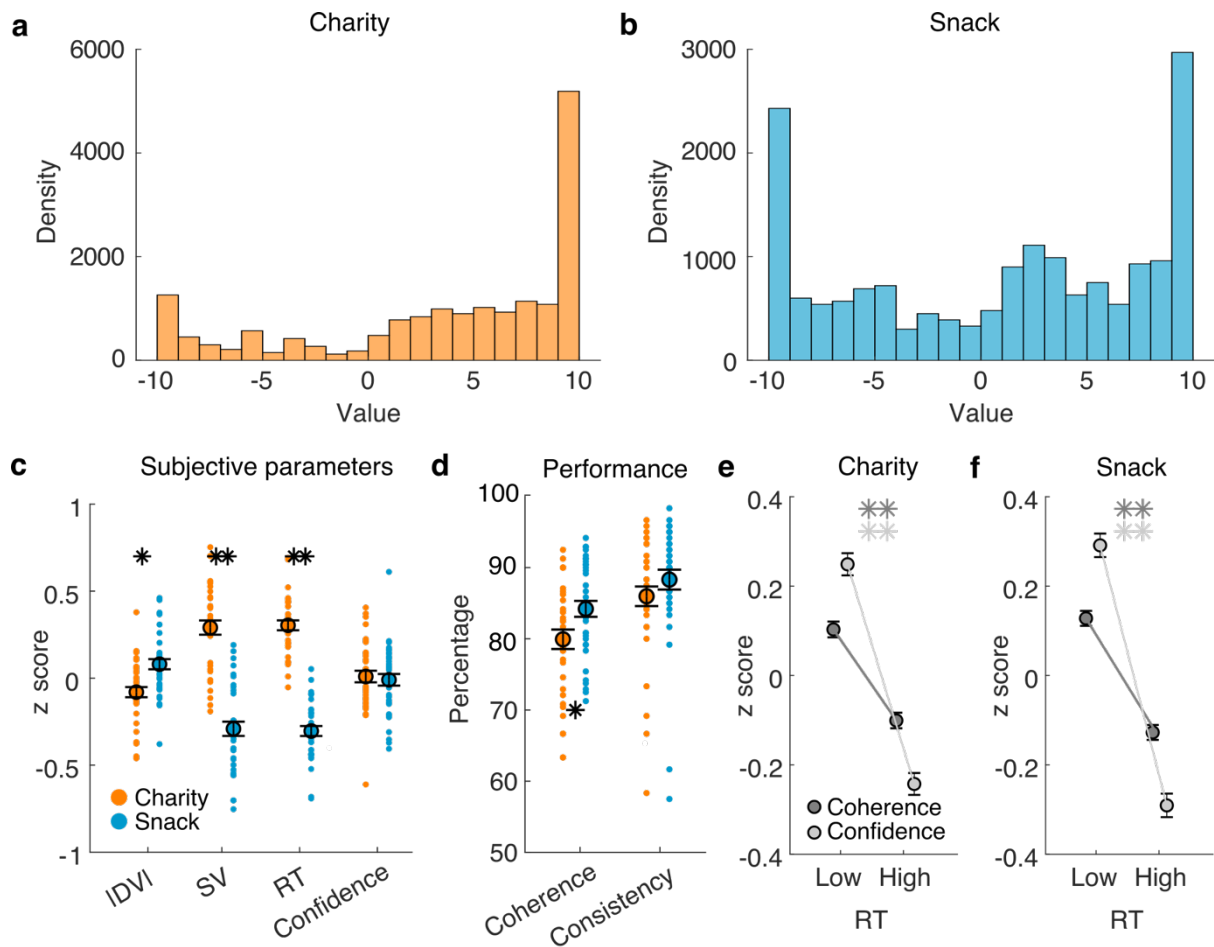
## References

- Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 6082–6087.  
<https://doi.org/10.1073/pnas.1800795115>
- Bhome, R., McWilliams, A., Huntley, J. D., Fleming, S. M., & Howard, R. J. (2019). Metacognition in functional cognitive disorder- a potential mechanism and treatment target. *Cognitive Neuropsychiatry*, *24*(5), 311–321.  
<https://doi.org/10.1080/13546805.2019.1651708>
- De Martino, B., Bobadilla-Suarez, S., Nouguchi, T., Sharot, T., & Love, B. C. (2017). Social information is integrated into value and confidence judgments according to its reliability. *Journal of Neuroscience*, *37*(25), 6066–6074.  
<https://doi.org/10.1523/JNEUROSCI.3880-16.2017>
- De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110.  
<https://doi.org/10.1038/nn.3279>
- Faivre, N., Pereira, M., Gardelle, V. De, & Vergnaud, J. (2019). *Confidence in perceptual decision-making is preserved in schizophrenia*. *MedRxiv*.  
<https://doi.org/10.1101/2019.12.15.19014969>
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), 17–19. <https://doi.org/10.1038/s41562-016-0002>
- Frankl, V. E. (1959). *Man's Search for Meaning: an introduction to logotherapy*. In *Language*. <https://doi.org/10.1080/10503300903527393>
- Hertz, U., Bahrami, B., & Keramati, M. (2018). Stochastic satisficing account of confidence in uncertain value-based decisions. *PLoS ONE*, *13*(4), 1–23.  
<https://doi.org/10.1371/journal.pone.0195399>
- Heyes, C., Bang, D., Shea, N., Frith, C. D., & Fleming, S. M. (2020). Knowing Ourselves Together: The Cultural Origins of Metacognition. *Trends in Cognitive Sciences*, *24*(5), 349–362. <https://doi.org/10.1016/j.tics.2020.02.007>
- Lysaker, P. H., Hillis, J., Leonhardt, B. L., Kukla, M., & Buck, K. D. (2014).

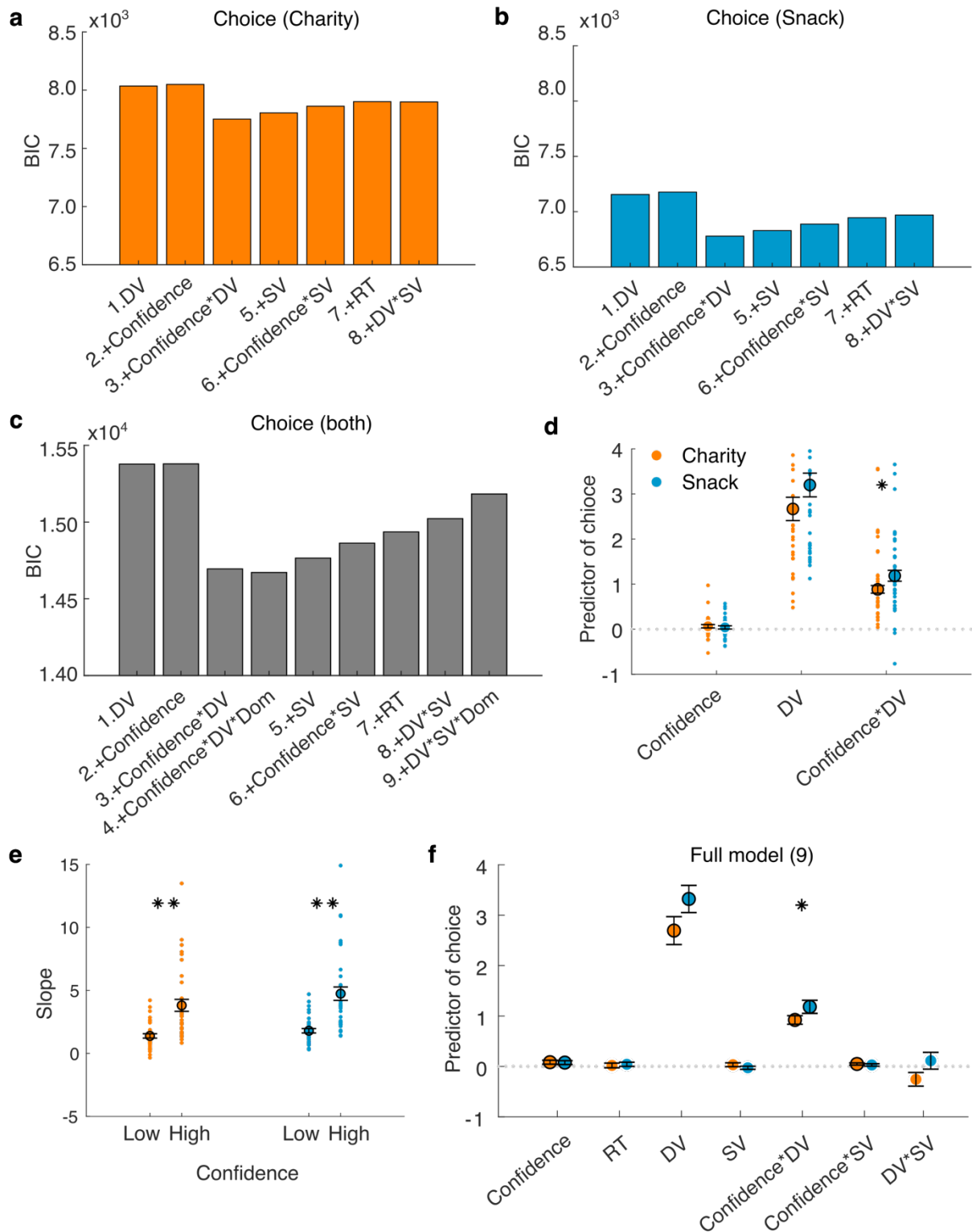


- Metacognition in Schizophrenia Spectrum Disorders: Methods of Assessment and Associations with Neurocognition, Symptoms, Cognitive Style and Function. *Isr J Psychiatry Relat Sci.*, 51(1), 54–62.  
<https://doi.org/10.1016/B978-0-12-405172-0.00006-5>
- Maoz, U., Yaffe, G., Koch, C., & Mudrik, L. (2019). Neural precursors of decisions that matter—an ERP study of deliberate and arbitrary choice. *ELife*, 8, 1–23.  
<https://doi.org/10.7554/eLife.39787>
- Odhnoff, J. (1965). On the Techniques of Optimizing and Satisficing. *The Swedish Journal of Economics*, 67(1), 24. <https://doi.org/10.2307/3439096>
- Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2(3), 511–527. <https://doi.org/10.1111/j.1756-8765.2010.01096.x>
- Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, 19(3), 366–374. <https://doi.org/10.1038/nn.4240>
- Rouault, M., McWilliams, A., Allen, M. G., & Fleming, S. M. (2018). Human Metacognition Across Domains: Insights from Individual Differences and Neuroimaging. *Personality Neuroscience*, 1(May).  
<https://doi.org/10.1017/pen.2018.16>
- Seligman, M. E., & Csikszentmihalyi, M. (2000). Positive psychology. An introduction. *The American Psychologist*. <https://doi.org/10.1037/0003-066X.55.1.5>
- Teodorescu, A. R., Moran, R., & Usher, M. (2016). Absolutely relative or relatively absolute: violations of value invariance in human decision making. *Psychonomic Bulletin and Review*, 23(1), 22–38.  
<https://doi.org/10.3758/s13423-015-0858-8>
- Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron*, 96(2), 348–354.e4.  
<https://doi.org/10.1016/j.neuron.2017.09.006>

## 6. Supplementary figures



**Fig S1: Value domains' behavioural features.** **a-b.** Distribution of subjective values for each item rated in the first part of the experiment (Fig 1a) on continuous scales from (-10) really dislike to obtain snack or support charity to (10) really like, respectively for charities and for snacks . **c.** Subjective parameters in both value domains being z scored within participants and across domains to observe differences between domain:  $|DV|$ = absolute difference in value, SV = sum of values, RT = response time. **d.** Differences between both value domains in norms of performance: coherence with one's subjective values and consistency (repeated decision)over the two presentations of each items pairs. **e-f.** Response time as implicit uncertainty predicting both confidence (explicit uncertainty) and choice coherence for both value domains (paired t-tests\*: $p < 0.05$ , \*\*:  $p < 0.001$ ).

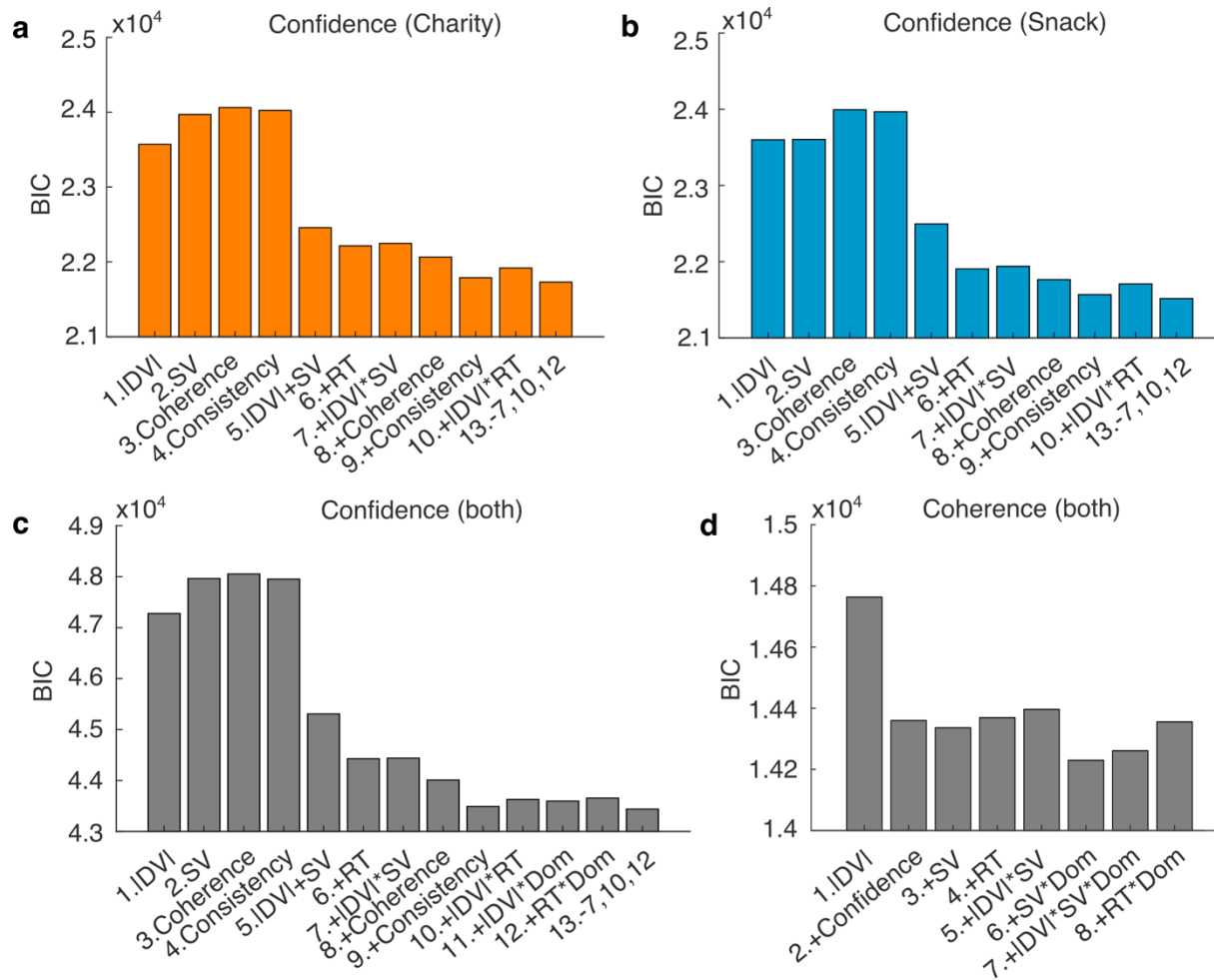


**Fig S2: Mixed models comparison for predictors choice.** a-c. Mixed models were compared for their BIC either in each value domain or with both to test effect of domain interaction. Significant estimates ( $p < 0.05$ ) are circles in black and error bar represents the SE of these estimates. The presence of a star represent the

explanation of a difference between both condition by an interaction term in the corresponding full model tested for comparison (\*Dom: \*=p<0.05, \*\*=p<0.001). **d.** Best model (number 4) of choice where the interaction between Confidence and DV is also explained by the difference in condition. Note that the emerging difference in difficulty between our two values domains in this design led to a better higher coherence in the Snack choices which might be linked with the difference in metacognitive insight between the two domains. **e.** Modulation of sensitivity by confidence in both value domains: for each participant the sensitivity slope for low and high confidence (Fig. 1c-d) is significantly different (paired t-test: \*\*=p<0.001). **f.** Model 9 with all predictors tested, highlighting their insignificance in predicting the choice.

#### Full models of Choice tested in Sup Fig. 2c:

1. choseRside ~ 1 + zDV + (1 + zDV | subj)
2. choseRside ~ 1 + zDV + zConfidence + (1 + zDV + zConfidence | subj)
3. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV + (1 + zDV + zConfidence + zConfidence\*zDV | subj)
4. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom | subj)
5. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV | subj)
6. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV | subj)
7. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT | subj)
8. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV | subj)
9. choseRside ~ 1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV\*zDom + (1 + zDV + zConfidence + zConfidence\*zDV\*zDom + zSV + zConfidence\*zSV + zRT + zDV\*zSV\*zDom | subj)



**Fig. S3: Mixed models comparison for choice confidence and choice coherence a-c** Comparison of models predicting choice confidence for either both values domains or both to test interaction of domain on predictors. (best model 12 is presented in Fig. 3b) **d.** Comparison of models predicting choice coherence as choosing the item with the highest value, model 7 is presented in Fig 3a.

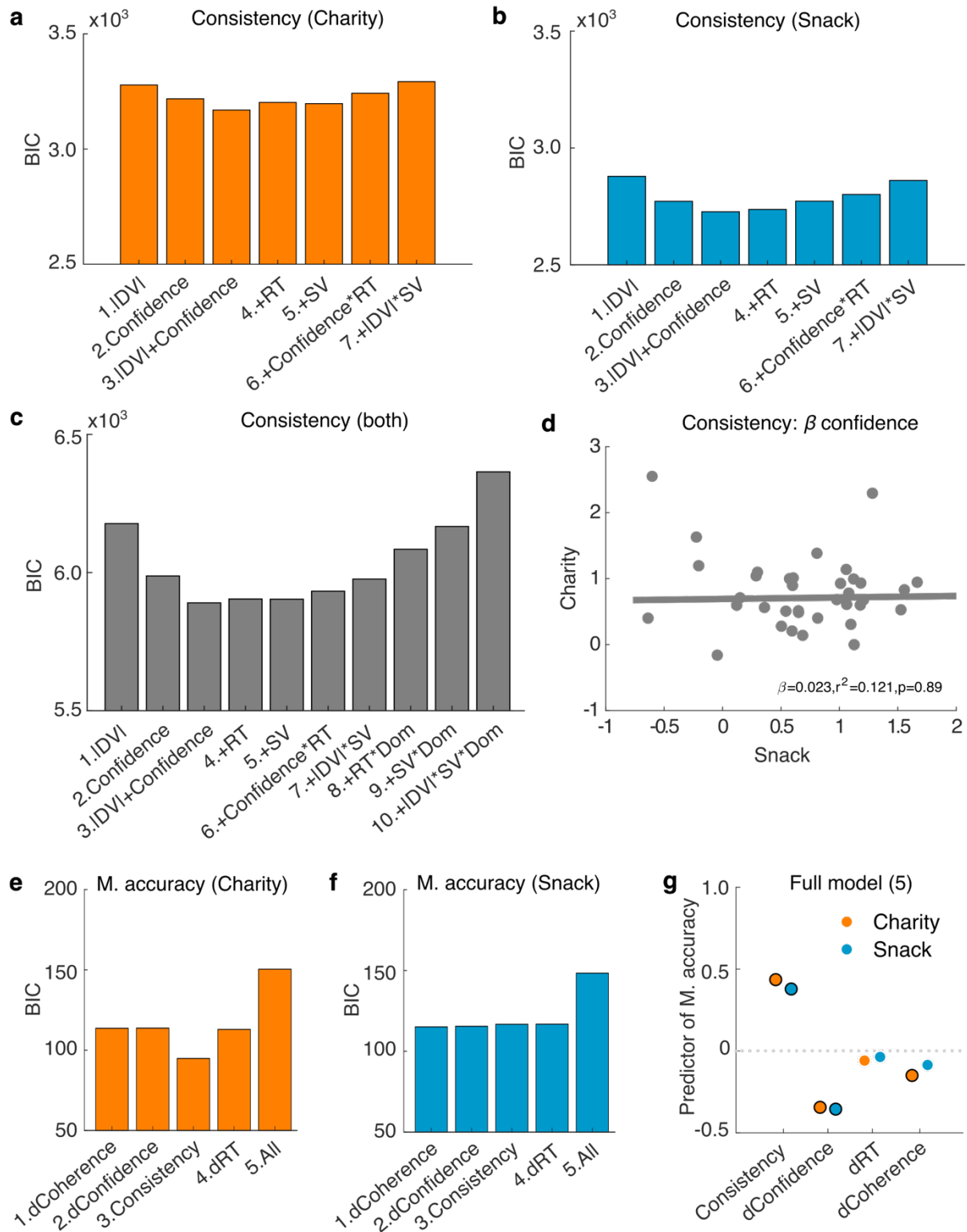
#### Full models of Choice coherence tested in Sup Fig. 3a:

1. Coherence ~ 1 + zaDV + (1 + zaDV | subj)
2. Coherence ~ 1 + zaDV + zConfidence + (1 + zaDV + zConfidence| subj)
3. Coherence ~ 1 + zaDV + zConfidence + zSV + (1 + zaDV + zConfidence + zSV| subj)
4. Coherence ~ 1 + zaDV + zConfidence + zSV + zRT + (1 + zaDV + zConfidence + zSV + zRT| subj)
5. Coherence ~ 1 + zaDV + zConfidence + zSV + zRT + zaDV\*zSV + (1 + zaDV + zConfidence + zSV + zRT + zaDV\*zSV| subj)
6. Coherence ~ 1 + zaDV + zConfidence + zSV\*zDom + zRT + zaDV\*zSV + (1 + zaDV + zConfidence + zSV\*zDom + zRT + zaDV\*zSV| subj)
7. Coherence ~ 1 + zaDV + zConfidence + zSV\*zDom + zRT + zaDV\*zSV\*zDom + (1 + zaDV + zConfidence + zSV\*zDom + zRT + zaDV\*zSV\*zDom| subj)
8. Coherence ~ 1 + zaDV + zConfidence + zSV\*zDom + zRT\*zDom + zaDV\*zSV\*zDom + (1 + zaDV + zConfidence + zSV\*zDom + zRT\*zDom + zaDV\*zSV\*zDom| subj)

#### Full models of Confidence tested in Sup Fig. 3c:

1. zConfidence ~ 1 + zaDV + (1 + zaDV| subj)
2. zConfidence ~ 1 + zSV + (1 + zSV| subj)
3. zConfidence ~ 1 + zCoherence + (1 + zCoherence| subj)
4. zConfidence ~ 1 + zConsistency + (1 + zConsistency| subj)

5.  $z\text{Confidence} \sim 1 + zaDV + zSV + (1 + zaDV + zSV | \text{subj})$
6.  $z\text{Confidence} \sim 1 + zaDV + zSV + zRT + (1 + zaDV + zSV + zRT | \text{subj})$
7.  $z\text{Confidence} \sim 1 + zaDV + zSV + zRT + zaDV*zSV + (1 + zaDV + zSV + zRT + zaDV*zSV | \text{subj})$
8.  $z\text{Confidence} \sim 1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence + (1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence | \text{subj})$
9.  $z\text{Confidence} \sim 1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence + zConsistency + (1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence + zConsistency | \text{subj})$
10.  $z\text{Confidence} \sim 1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence + zConsistency + zTransitivity + (1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence + zConsistency + zTransitivity | \text{subj})$
11.  $z\text{Confidence} \sim 1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence + zConsistency + zTransitivity + zaDV*zRT + (1 + zaDV + zSV + zRT + zaDV*zSV + zCoherence + zConsistency + zTransitivity + zaDV*zRT | \text{subj})$
12.  $z\text{Confidence} \sim 1 + zaDV*zDom + zSV + zRT + zaDV*zSV + zCoherence + zConsistency + zTransitivity + zaDV*zRT + (1 + zaDV*zDom + zSV + zRT + zaDV*zSV + zCoherence + zConsistency + zTransitivity + zaDV*zRT | \text{subj})$
13.  $z\text{Confidence} \sim 1 + zaDV*zDom + zSV + zRT*zDom + zaDV*zSV + zCoherence + zConsistency + zTransitivity + zaDV*zRT + (1 + zaDV*zDom + zSV + zRT*zDom + zaDV*zSV + zCoherence + zConsistency + zTransitivity + zaDV*zRT | \text{subj})$
14.  $z\text{Confidence} \sim 1 + zaDV*zDom + zSV + zRT + zCoherence + zConsistency + (1 + zaDV*zDom + zSV + zRT + zCoherence + zConsistency | \text{subj})$



**Fig S4 Mixed model comparison for choice consistency and subjective metacognitive accuracy.** **a-c.** mixed models comparison for choice consistency in second presentation of the choices with confidence at the previous encounter of the item pair, respectively for each value domains and both to test for domain interaction, model 5 is presented in Fig. 2a. **c.** Linear regression for both factor of

confidence as predictor of choice consistency between both domain across individuals. **e-f.** Mixed model comparison for participant metacognitive ability as measured from best model in Fig S2d comparing individual parameters as candidate predictor of this subjective insight and all parameters combined in model 5. **g.** Fixed effects of model 5 as predictors of metacognitive accuracy in both domains for behavioural differences between first and second presentation of each choices, to study behavioural optimisation over time: Consistency: binomial measure if same item is selected twice, difference variables between the second and first presentation of a pair of items: difference in confidence (dConfidence), difference in response time (dRT), difference in binary Coherence (dCoherence).

Full models of Choice consistency over the 2 presentations tested in Sup Fig. 4a:

1. Consistency ~ 1 + zaDV + (1 + zaDV | subj)
2. Consistency ~ 1 + zConfidence + (1 + zConfidence | subj)
3. Consistency ~ 1 + zaDV + zConfidence + (1 + zaDV + zConfidence | subj)
4. Consistency ~ 1 + zaDV + zConfidence + zRT + (1 + zaDV + zConfidence + zRT | subj)
5. Consistency ~ 1 + zaDV + zConfidence + zRT + zSV + (1 + zaDV + zConfidence + zRT + zSV | subj)
6. Consistency ~ 1 + zaDV + zConfidence + zRT + zSV + zConfidence\*zRT + (1 + zaDV + zConfidence + zRT + zSV + zConfidence\*zRT | subj)
7. Consistency ~ 1 + zaDV + zConfidence + zRT + zSV + zConfidence\*zRT + zaDV\*zSV + (1 + zaDV + zConfidence + zRT + zSV + zConfidence\*zRT + zaDV\*zSV | subj)
8. Consistency ~ 1 + zaDV + zConfidence + zRT\*zDom + zSV + zConfidence\*zRT + zaDV\*zSV + (1 + zaDV + zConfidence + zRT\*zDom + zSV + zConfidence\*zRT + zaDV\*zSV | subj)
9. Consistency ~ 1 + zaDV + zConfidence + zRT\*zDom + zSV\*zDom + zConfidence\*zRT + zaDV\*zSV + (1 + zaDV + zConfidence + zRT\*zDom + zSV\*zDom + zConfidence\*zRT + zaDV\*zSV | subj)
10. Consistency ~ 1 + zaDV + zConfidence + zRT\*zDom + zSV\*zDom + zConfidence\*zRT + zaDV\*zSV\*zDom + (1 + zaDV + zConfidence + zRT\*zDom + zSV\*zDom + zConfidence\*zRT + zaDV\*zSV\*zDom | subj)

Full models of subjective metacognitive accuracy tested in Sup Fig. 5:

1. metaAcc ~ 1 + dCoherence + (1 + dCoherence | subj)
2. metaAcc ~ 1 + dConfidence + (1 + dConfidence | subj)
3. metaAcc ~ 1 + Consistency + (1 + Consistency | subj)
4. metaAcc ~ 1 + dTransitivity + (1 + dTransitivity | subj)
5. metaAcc ~ 1 + dRT + (1 + dRT | subj)
6. metaAcc ~ 1 + dCoherence + dConfidence + Consistency + dRT + dTransitivity + (1 + dCoherence + dConfidence + Consistency + dRT + dTransitivity | subj)



# Chapter 7

## General Discussion

This thesis aims at highlighting the key position of subjective value (*i.e.* learned cost-benefit outcome of a decision) in building a comprehensive picture of procedural metacognition (*i.e.* monitoring and control). From an ever-present integration of value to inform monitoring signals (*e.g.* confidence) to the central position these take in ensuring the reliability of decisions reliability coherence (*e.g.* coherence between goal and decision), we argue that subjective value is a key bridge between both these sides of procedural metacognition. While in cognitive neuroscience most of the current literature on metacognition relates to perceptual decisions, we argue that these reliable but simple tasks limit the insight we can gain in light of the complexity of human metacognition (*c.f.* Chapter 2).

Assuming that in humans explicit metacognition is developed up to adulthood to make autonomous and responsible member of society (Bigenwald & Chambon, 2019) we attempt to account for models of metacognition that might describe up to complex life-like decisions. As in these tasks there is often no such thing as simply right and wrong answers, we take the perspective that metacognition monitors and optimise cognition and behaviour by aiming at a desired cost-benefit ratio we refer to as desired reliability (Ackerman & Thompson, 2017; Shea & Frith, 2019). We suggest that, throughout increasingly complex tasks and executive controls, metacognitive monitoring relies essentially on an ubiquitous monitoring of subjective value (Fig 1). Throughout the thesis, we presented conceptual and empirical work to argue for and test the hypothesis that (throughout a range of tasks) subjective value is ubiquitously computed (throughout different tasks) in monitoring signals to inform the agent about the reliability of her decisions. Following the notion of procedural metacognition that is embedded in bounded rationality, we suggest that this account of subjective value in monitoring signals is key to understand metacognition as a thermostat that ensures a desired level of decision reliability by fitting effort involved in the decisions process.

The above proposition aims at filling the following gap: On the one hand, the concept of procedural metacognition suggests that this higher order enables agents to monitor and adjust their cognition and behaviour to increase their probability of success and eventually reach their goals. However in science, both these function of subjective monitoring of success and of control of one's cognition and behaviour are mostly studied independently in human subjects. The monitoring side that mainly aims at defining how explicit monitoring signals (e.g. feeling of knowing, confidence ..) are formed is studied in the field of metacognition whereas the control side that looks at how one regulates effort and performance is studied by the field of executive control. On the other hand, in the field of metacognition (not so procedural anymore), certain questions about the nature of metacognition (such as in regards to its generality across tasks or its inefficiencies in tracking accuracy) highlight possible gaps in the current models used to define the computation of metacognitive monitoring signals. Essentially, the growing research on metacognition that incorporates subjective value as input or that suggests it presents a hierarchical structure appear as promising keys to unlock a more comprehensive view of the working of metacognition as a whole (Shea & Frith, 2019; Soutschek et al., 2021). Here we propose to focus on the former by integrating subjective value in the models that explain how metacognitive signals are formed, in the light of serving a function of control.

While recent studies aimed at joining both fields by relying on behavioural economics to operationalise complex behavioural tasks, in this thesis we ask the following question:

**What does accounting for the subjective value of decisions contribute to our understanding of the function and computation of metacognition?**

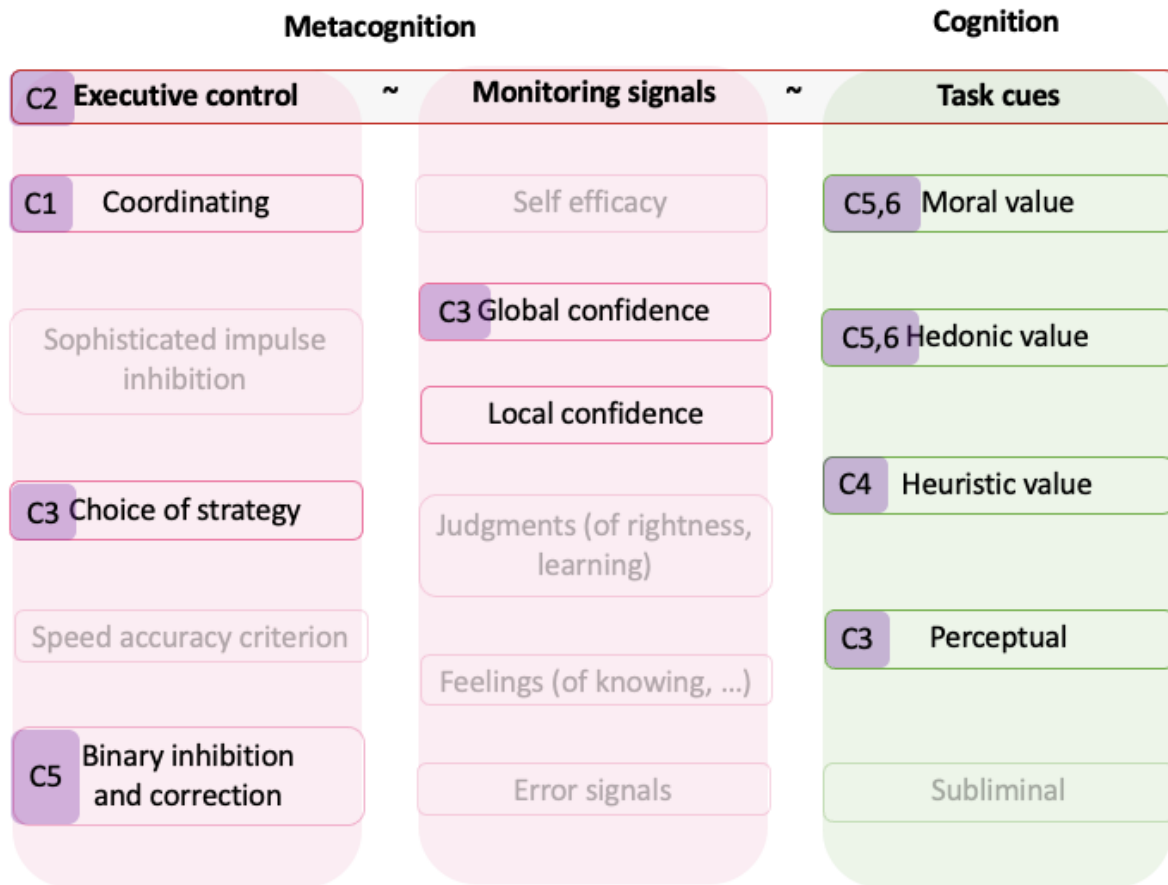
Throughout the various (conceptual and empirical) chapters of the thesis, we thrive to both propose and test a framework for procedural metacognition which essentially acts as a thermostat for decision reliability by its access to subjective value. While we define the concept of decision reliability in a bounded rationality framework, at the level of local decision, we use the proxy norm of coherence

between the decision and subjective value. Together, our analogy of metacognition as a thermostat for decision reliability can be reduced in a parsimonious bidimensional map of procedural metacognition (Fig. 1): on one dimension the processing from input to monitoring signal to eventual control, on the other dimension, an array of elements fit for decision tasks of various complexity. We thereby suggest that while subjective value might be discarded when studying the most simple of decisions, it quickly becomes essential to account for when studying more complex decisions, mainly where the decision rule to follow is not clearly instructed but relies on subjective experience and intentions.

In this general discussion, we will start by summarising the main contributions of all chapters in this proposed framework which aims at illustrating the place of subjective value in procedural metacognition. In a second part, we will discuss more broadly how each of these findings build together a more comprehensive picture of metacognition by adding subjective value as a central input to its monitoring. Lastly, a third part will propose a broader contribution of the present picture of metacognition both for research perspectives and beyond, in the real world.

### **1. Our findings support a value-based model of procedural metacognition**

The present thesis aimed at closing the loop between the computation and function of procedural metacognition from the simplest to more complex tasks. To do so, we claim that subjective value plays a critical role in enabling metacognition to act ubiquitously as a reliability thermostat. While the notion of reliability here is anchored in bounded rationality by considering the effort involved in the decision process, putting aside the control facet of metacognition, the monitoring of reliability can be simplified by two norms: globally the maximisation of the goal when choosing decision rule (in learning tasks); and locally the maximisation of the decision rule when the choosing of the item (i.e. coherence in preference tasks).



**Figure 1: Chapters contribution to the role of subjective value in procedural metacognition.** Procedural metacognition suggests that executive control functions (left) are modulated by metacognitive monitoring signals (middle), themselves modulated by available cues (right). As developed in Chapter 2, we suggest that throughout a not-so-linear evolution of species and development of human agents, the maturation of explicit awareness supports the apparition of new abilities to control behaviour and become aware of oneself and one’s values. The different chapters (purple tags) aim at exploring the contribution of subjective value to this both-sided picture of procedural metacognition. The first three chapters are conceptual and setting this framework. Chapter 1 suggest that some of humans’ unique skills (as to being a moral hero), rely essentially on the monitoring of the coherence between one’s decisions and one’s (moral) values. Chapter 2 discusses in philosophy and in cognitive neuroscience the place of monitoring signals in procedural metacognition working as a thermostat for decision reliability. Chapter 3 proposes a model linking value and learning to perceptual metacognition to articulate the ubiquitous link between value and metacognition in adjusting

strategies. The second part with three empirical chapters aim at testing the model from the cognitive neuroscience perspective: with the computation of local confidence signals in regards to their cues. Chapter 4 explores the contribution of heuristics to both the reflective functions of metacognition and theory of mind suggesting they both rely on the decision rule (or strategy applied). Chapter 5 demonstrates the access of moral values by metacognition and suggests that these local confidence signals might be linked to the consistency with which these decisions are repeated over time. Chapter 6 suggests that across value domains, metacognition does not appear to necessarily track whether a choice was optimal (relative to the set of options) but rather of sufficient value (independently from alternative options) supporting the view that metacognition regulates behaviour as a thermostat to reach a desired criterion of cost-benefit. Chapter 7 discusses the implications of accounting for subjective value in procedural metacognition.

Both first chapters of the thesis aimed at objectively stating the art of academic research in metacognition. These chapters come to define procedural metacognition as a thermostat that allocates the required amount of effort to a task in order to reach a desired level of reliability between one's decisions and goals. While we highlighted the current gap in research to address metacognitive monitoring signals as themselves cues on which executive control might rely, we briefly mentioned the recent merging of the field of economics with research in metacognition which seems very promising to address this issue (Lieder et al., 2018; Soutschek et al., 2021).

In Chapter 3 we review the literature to propose a novel computational model of metacognition where both perceptual and value-based cues are ever-present components of confidence reports serving different informative roles. This suggestion that monitoring signals are ever-present (even implicitly) to track the fit between decisions and contexts is widely agreed in the field (Pasquali et al., 2010; Shea et al., 2014; Shea & Frith, 2019). However how these different monitoring signals are computed remains unclear. We suggest here that by accounting for their function as regulator of cognition and behaviour, we might get better insight in the computational (and eventually neurobiological) roots that define them.

Chapter 4 studies the explicit access of metacognition to heuristic cues. While it was previously demonstrated that confidence signals went hand in hand with popular opinions, we wondered whether participants could be conscious of such popular trend and thereby influence their metacognitive monitoring by their theory of mind. Our findings indeed reveal that both reflective systems appear to rely on the same cues revealing that, in tasks with limited knowledge, both this reflective system use the same decision rule to reflect upon their choices and others. This finding relates to the previous chapter (3) highlighting the nature of reflective system as evaluating a decision (of oneself or others) in light of a decision rule as a hierarchical comparison between a global decision rule and a local action (Rouault et al., 2019). Chapter 5 aimed at testing whether metacognition was able to monitor and eventually guide our moral decisions. While relevant on its own as a research question, the moral value domain was selected also for the specific place it is believed to hold in conscious and voluntary decisions, as opposed to hedonic decisions which, by their reward might be more habitual than goal-directed (Maoz et al., 2019). Our results thereby suggest that metacognition has access to a wide range of value domains and might always rely on subjective value to guide decisions in a given context, whether is learned by direct feedback or relying on higher cognitive skills such as theory of mind (Kitcher, 2011).

Chapter 6 finally looks at metacognition of value-based decision across domains with preliminary data. The thesis suggesting that subjective values are essential to metacognition which serves at adjusting bounded rationality, we aimed at testing whether metacognitive monitoring appeared rather optimal or satisfiable in nature. We found significant evidence that participant might, across value domains, share a similar criterion for tracking either optimally or satisfiable. These results support the idea of an ever present monitoring of subjective value by metacognition but also eventually to do so with a domain-general cognitive system setting similar evaluative standard across types of tasks. Furthermore, suggesting the possibility of a subjective level of satisfiability across tasks, the results contribute to the discussion of mental health relating to such a subjective tendency to monitor oneself at the subjective level (Seow et al., 2021). Further than relating to normative questions (i.e. accuracy and bias) as previous literature focussing on perceptual research has done,

these results propose different metacognitive monitoring styles (Rouault et al., 2018). These preliminary results therefore support the view that the computation of value is central to metacognition which is embedded in bounded rationality.

Altogether our results suggest an ever-present monitoring of subjective value across different domains and that, depending on the level of conscious awareness, agents may have associated abilities to exert control over their environment (Hohwy, 2013).

## **2. Remaining questions for this value-based model of procedural metacognition.**

We presented a model which suggests that value-based input are ubiquitous in the computation of metacognitive signals: from error signals associated with reflexive corrections to explicit appraisals when proactive action plans are required. However such a broad picture opens up many areas for research beyond what has been attempted in the present work. Focussing on empirical work, we now discuss some of the limitations of our empirical studies and some remaining questions to test or further assert this framework. In a second part, we then focus more broadly on the field of metacognition in cognitive neuroscience and touch upon how the present framework could contribute to central gaps in the field.

### **2.1. Research limitations and directions**

#### *Operationalising value in the laboratory*

Empirically, we aimed at testing the ubiquitous contribution of value (across various domains) to the computation of metacognitive signals (i.e. confidence levels) and their sensitivity. To do so we designed a novel paradigm that enables to study moral preferences similarly to hedonic preferences. Nonetheless, our experiments could be seen as limited. First, the paradigm of revealing preferences by choices was initially designed by incentivising participants with their decisions affecting the probability of receiving a reward at the end of the experiment (De Martino et al., 2013; Folke et al., 2017). Future experiments, offering to link moral choices to donations or hedonic choices to obtention of snacks could eventually motivate

participants to take the task more seriously and strengthen the performances and sensitivities obtained here.

Secondly, regarding the question of domain generality, we have aimed at endorsing the challenge in two different manners: either by asking participant to report their confidence in across-domains choices (a charity and a snack) or by asking them a comparative confidence between two within-domain choice they were more confident about. In the first instance, we found that choices were not predicted by reported subjective values and therefore could not test a metacognitive ability, in the second we found no sensitivity of the comparative confidence to the difference of value between pairs of items. In other words, both our experiments rendered us without conditions to conclude anything about domain generality of confidence across value domains. However, a different paradigm with an economic wager as common currency instead of a choice followed by a confidence report appear to be a useful vector to get participants to perform in accordance to their reported subjective values and trade between reward for self or other (Moll et al., 2006). We believe that this choice paradigm could be a promising direction to undertake this question.

### *Subjective traits*

Our last chapter aimed at testing a possible subjective criterion of optimality in across value domains as participants sensitivity to the relative value of their choice as opposed to its absolute value. While this study aimed at finding across metacognitive sensitivity some evidence for a subjective criterion for value based task across domains, our findings though significant should require larger number of participants to validate this study. Furthermore, the study of a common metacognitive bias across value domains would greatly contribute to explaining the role of metacognition in psychological disorders such as anxiety and depression as this link has only been reported in simple perceptual tasks(Rouault et al., 2018; Seow et al., 2021).



*Types of value*

Lastly, our research also aimed at testing how contextual knowledge affected metacognitive ability. While the present experiment (Chapter 4) aimed at testing the link between metacognition and theory of minds as similar reflective processes, we suggest that beyond task knowledge, self-knowledge might also play a role in metacognitive sensitivity. More particularly, whether in moral decisions a stronger sense of identity could be linked with a greater metacognitive ability. While it was demonstrated that agents with extreme political views tended to have overall lower metacognitive ability (in perceptual tasks) (Schulz et al., 2021), one could ask whether people with a strong sense of identity (independently of their extreme tendency) could have stronger metacognitive abilities in value-based, compared to participants with a lower sense of identity.

**2.2. Contribution to metacognition**

In cognitive neuroscience, recent efforts were made to agree on common goals to be met in order to advance the understanding of the computation of confidence reports (Fleming, 2023; Rahnev et al., 2021). We suggest that research focussing on tasks that merge both control and monitoring will provide profound insight in the field by considering explicit reports as these intermediate signals between both processes. Mainly, the interaction of behavioural economics and metacognition could be particularly fruitful to explain in both field how certain complex behaviour emerge such as reasoning (Ackerman & Thompson, 2017; Lieder et al., 2018), choice of strategy or sophisticated impulse inhibition (Soutschek et al., 2021).

Nonetheless our model of subjective value being central to the procedural function of procedural metacognition could be also questioned. Indeed, metacognitive blindsight as in choice blindness (Hall et al., 2010) or gap between confidence levels and corrective behaviour (Vaghi et al., 2017) suggest that explicit metacognition could have roles that are not essentially procedural, but maybe of communicating consistency (Koriat & Adiv, 2015). Furthermore, we suggested throughout the thesis that explicit metacognition is important to both navigate social norms efficiently (as an autonomous agent) but also to build a set of subjective values that are personal, building a sense of identity that can detach from social norms (e.g. maverick). While

the roles of metacognitive ability in learning to adapt efficiently were empirically tested (Roebbers, 2017), the more longitudinal role of metacognitive development on the autonomy of agents remains to be implemented such as with an school curriculum (or cultural study). Recent efforts are nonetheless made to provide metacognitive enhancement tools an boost decision performance.

### **3. Open discussion on place and implications of our model**

In this last part, we aim to move beyond academic research and open the broader implication of incorporating subjective value in the study of procedural metacognition. Therefore, the following ideas are more of speculative thoughts to ponder upon than claims that are necessarily backed up by science or even testable as such.

Let's observe that, throughout the universe, sustainable and lasting systems are often maintained by regulating mechanisms: whether it be counter-balancing forces, feedback loops or more hierarchical monitor-and-control mechanisms, the presence of such a flexible adjustment mechanism enables the systems to pervade throughout time and ever-present change. In the present thesis, we introduced the idea of metacognition of subjective value being similar to a thermostat that regulates the amount of effort to be allocated in order to maintain a certain degree of coherence between one's intentions and decisions. Together with this backbone for coherence, we suggested that it also supports an agent's ability to remain autonomous and detach from collective behaviours and heuristics. Here, we wish to elaborate on such associated traits and ask:

#### **What abilities might human gain by having their metacognition access their subjective values?**

To explore the driving forces behind the link between metacognition and value, we will examine three distinct levels of analysis: first, the social level, followed by the individual level, and finally the cognitive level. We will then return to discussing potential research directions for value-based metacognition. First, at the social level, we will ask about the advantages that might be provided to an agent whose decisions remain coherent with her intentions. Secondly, at the individual level, we will touch upon theories that discuss how this coherence, at different levels, might

enable the agent to reach a healthier psychological state. Thirdly, cognitively, we will question the cognitive markers that might support reaching these levels of self-coherence.

### **3.1. Social advantages of self-coherence**

#### **3.1.1. Self-definition**

What selective forces might be at play to nudge humans to be coherent? What would it mean to be coherent for such a social function?

We defined metacognition of value as the ability to allocate effort appropriately in order to overwrite habitual, heuristic or instinctive behaviour and instead engage into more demanding counter-intuitive, reflexive, and challenging behaviour when it is required in order to remain coherent with one's intentions. By accounting for social (e.g. conceptual) value, human metacognition might therefore support the agent's ability to self-define its cognition and behaviour in regard to a social environment and its norms. In this sense, can we say that metacognition acts as a catalyst for achieving higher levels of coherence in such complex environments?

Coherence can manifest at various levels depending on the subjective values involved. We define subjective value as the learned cost or benefit associated with an action (whether conscious or not). According to Free Energy Theory, living beings continuously learn these values to regulate their behaviour effectively. Building on this, Timmermans (et al., 2012) proposes an evolutionary theory that suggests that animal species might have developed up to three such levels of regulation: regulating internal processes, regulating interactions with the world, and regulating social interactions with other agents. In this framework, metacognition extends beyond ensuring coherence with the natural environment (e.g., recognizing the value of a nutritious berry versus a poisonous one) to also supporting coherence within social groups (e.g., weighing the value of stealing a berry from an alpha).

Humans stand out due to their deeply ingrained capacity for engaging with a rich social environment from infancy. This unique trait enables humans not only to share a cognitive space for coordinating collective goals (e.g., cooperative hunting) but also to share a collective mind: a cultural environment with social norms, knowledge and institutions that are just as real to humans as the natural environment is to other

species (Tomasello et al., 2012). Thus, we argue that human metacognition, grounded in conceptual language, plays a crucial role in helping individuals navigate and integrate into this complex cultural system. Depending on the species and the associated nature of subjective values, we suggest that metacognition helps ensure that agents maintain coherence by aligning their actions with their intentions whether relating to natural or cultural environments.

At a second level, we suggest that human metacognition can support the development of the agent's identity. We call self-definition the formation of such identity by endorsing a collection of conceptual beliefs and intentions, that might themselves come with a given degree of alignments between them. While metacognition is known to develop tardively with the prefrontal cortex up to young adulthood (when one becomes as a rational and responsible autonomous citizen), here we suggest that this development of metacognition at this age goes hand in hand with the development of the agent's identity. While it was demonstrated that low metacognitive ability is linked to extreme political views, we propose that metacognitive ability might itself foster the emergence of complex identities more finely tuned to the structure of the social environment therefore less likely to detach from extreme social norms. In other words metacognition could develop the coherence of the agent social status within the social structure by finding a role that increase coherence within both systems and their interaction.

Lastly, this ability to integrate in a cultural environment might contribute to the coherent structure of the social environment through their defined role. This contribution of the self-defined agent to the cultural structure might itself come in different degree depending on how much the agent manages to act in accordance with her value. For example, creativity could be a form of self-expression that reflects the good coherence between identity and cultural environment. At a higher level of coherence, a moral hero could go against social norms to act according to their beliefs. At the highest level, an individual might lead a group to reform the cultural structure itself, thereby expending the levels of coherence from his self-defined identity (beliefs) to increase the coherence of his environment.

To summarise, we suggest that, different animal species, relying on different complexity of subjective values, might have different levels of complexity when it comes to coherence between their values and decisions. For humans, the conceptual and cultural values might lead metacognition to develop the agent's identity which, itself, can come with different levels of coherence. If the identity coherence is higher than the environmental coherence, metacognition could support actions taken despite high risks or negative feedback to change the environment. We suggest that when optimally developed, metacognition supports higher levels of autonomy (see section 3.2) and now discuss some forces of selection that might have supported this role of metacognition as regulating coherence within the social environment.

### 3.1.2. Differentiation

Following theories on the roles of explicit metacognition, we take the stance that human metacognition might benefit the group by enabling agents to form a sense of identity, and eventually act accordingly to contribute to the social group.

Similar to cell differentiation in biology, the term here refers to functional specialization, such as adopting a specific role within the social group. While the terms self-definition or identity-formation focus on the agent's development, the term differentiation suggests the endorsement of a defined social role within the group, becoming part of (or eventually implementing) its given structure and complexity. While levels of distinction from the existing social structure will be discussed in part 3.2, here we simply discuss the implication of differentiated social roles within groups.

Most democratic countries provide education for all to elevate and self-regulate the nation. Beyond political philosophy's rich literature on what "ought to be" in terms of ethics and economics, one might ask whether the differentiation of its units truly elevates a system. For instance, do animals with more complex cell differentiation than plants have any advantage over them? Does the complexity of birds' biology benefit them over worms? In essence, the theory is that systems composed of specialised sub-units possess greater adaptive power than those composed of

conforming ones (Luhmann, 1977). Simply put, complexity breeds adaptability and sustainability of the system.

While metacognition of decision making is often seen as an advantageous feedback loop for improving agent performance, we suggest that applied to a social structure, it could refine the agent's identity to find a well-fitted social role. Thus, metacognition could act as a feedback loop on the social structure by supporting the human ability to complex cultural environment and, in turn, contributing to its complexity and adaptative power. By accessing conceptual and cultural value, human metacognition might offer evolutionary and developmental advantage that support coherence and complexity within individuals and their groups.

### 3.1.3. Competition

Evolutionary theories suggest that resource scarcity pressures social species to select the members within their groups (Tomasello et al., 2012). Economics models of decision theory demonstrate that, even non-human animals do prefer to interact with a trustworthy or skilled partner, bringing up the concept of reputation as selective pressure. Studies on group decision-making show that individuals prefer advice from those who can accurately communicate their confidence rather than from individuals with mere high confidence.

Thus, while metacognitive ability improves an agent's performance, it also enhances the agent's ability to gain connections and status within a social group. This social position, in turn, offers evolutionary benefits, favouring the development of explicit metacognition as a functional advantage in cooperation over less fortunate social groups. We have discussed how metacognition, incorporating social value, develops alongside an agent's social identity and contributes to group dynamics. In the following section, we will delve deeper into the function of metacognition in regulating agents' actions regarding social norms, defining incremental levels of this self-coherence.

### 3.2. Levels of self-coherence

Different theories address the human “need” to achieve high levels of coherence in their identity as a cornerstone of psychological development. While some, such as Pouget’s “True Self”, focus on overcoming traumas to free the agents from limiting thought patterns, this section explores frameworks in positive psychology that emphasize the pursuit of psychological development to its fullest potential. Among these, Abraham Maslow’s theory of self-actualization proposes a hierarchy of needs culminating in personal fulfilment. Similarly, Carl Jung’s process of individuation involves integrating the subconscious with the conscious mind to form a unified and complete psyche. Friedrich Nietzsche’s concept of the *übermensch* (or "overman") embodies the ideal of achieving full potential by transcending societal norms and creating independent values.

These theories converge on the idea that higher levels of coherence within one’s identity often manifest through creativity. We propose that this drive to create stems from a perceived mismatch between the coherence of one’s inner world and the external environment. An individual with a coherent worldview may see gaps or inconsistencies in their environment (whether social or natural) and feel compelled to bridge this difference in coherence through creative expression. Conversely, a lack of inner understanding may motivate an agent to create as a means of unveiling this inner incoherence in the world. In essence, we propose that the type of creativity that converges in many models of psychological fulfilment stems from an acute metacognitive skill that not only fosters a sense of self, but also drives the agent to create in order to align their inner coherence with their external environment. Secondly, numerous theories of cognitive development (e.g. Piaget, Erikson, Kohlberg, Montessori, Kegan) consider psychological development from infancy to adulthood to be driven in part by a shift in perspective. There the prefrontal cortex matures to account for an ever-more complex model of the world, shifting the agent’s perspective from a very subjective egocentric perspective to an ever more objective allocentric one. Putting these two together, we propose that these higher metacognitive skills in highly developed psyche provide these humans with higher levels of autonomy.

Previously, we argued that metacognition—the ability to monitor and adjust cognition and behaviour—is central to the concept of autonomy. Drawing on evolutionary cognition models that highlight the emergence of species with increasing capacity to solve complex tasks, we posited that value-based metacognition underpins the autonomy necessary for social cooperation. Tomasello’s framework of autonomy illustrates this progression across species on a timeline ranging over 300 million years as follow:

1. **Reactive Autonomy:** Found in lizards, involving choices based on environmental stimuli.
2. **Motivational Autonomy:** Observed in mammals, where inner models guide goal-directed behavior based on emotions and motivations.
3. **Rational Autonomy:** Evident in apes, characterized by flexible planning, tool use, and understanding others’ intentions.
4. **Normative Autonomy:** Unique to humans, driven by conceptual language and cultural norms, enabling cooperation and group coordination.

Beyond the latter evolution of normative autonomy, we propose that human metacognition, being explicit and conceptual, provides the cognitive foundation for creativity and innovation. This now developmental advantage provides human with the opportunity to adopt, within their lifetime, the metacognitive skills necessary to achieve higher levels of autonomy, integrating levels of coherence of increasingly complex spheres that transcend social norms. Just as the evolution of a central nervous system with neurotransmitters provided animals with a faster, more efficient mechanism for integrating signals compared to hormones, the development of social norms, language, and culture endowed humans with a prefrontal cortex as a new medium capable of accessing entirely new realms of coherence beyond the immediate natural environment.

If nurtured through the agent’s development, this prefrontal cortex can both tap into higher realms of the environment and provide metacognitive access to them. While evolutionary levels 1 to 4 enabled agents to fit in their environment, these levels of psychological developments transform agents into vectors for enhancing the coherence and sustainability of these broader spheres. Factors such as executive



control, resources, mentorship, and mental health increase the likelihood that these potentials will be realized through effective action. In this sense, the prefrontal cortex, housing human metacognition, becomes a powerful medium for connecting with higher spheres and other minds in these realms. This interaction forms cultural networks that refine complex and differentiated individuals (see section 3.1.2). While metacognition is not strictly necessary for agents to be interested in or connect with these higher spheres, it facilitates the creation of value by fostering awareness of coherence within and between these spheres. Ultimately, creating such value (often for the “greater good”) enables individuals to achieve higher states of psychological development, extending their identity’s coherence into legacies that transcend them by fulfilling their metacognitive potential.

We suggest three additional levels of autonomy that extend Tomasello’s framework, each grounded in increasingly sophisticated metacognitive skills and reflective of specific archetypes: the Maverick, the Hero, the Leader and the Alchemist.

### **5. Reflective Autonomy: the Maverick**

Reflective autonomy is the benchmark of higher levels of autonomy where metacognitive skill go beyond fitting the agent’s behaviour to the surrounding environment, whether natural or social. This allocentric perspective sees the individual not as central to the environment but as part of it. Removing this limiting reference point enables them to critically question and reframe existing norms. We suggest that this questioning fosters a stronger and more coherent sense of identity, granting the autonomy to transgress social norms in the pursuit of authenticity and creativity.

Nietzsche’s *übermensch* epitomizes this archetype, as does Truman from *The Truman Show*, who rejects the constructs of their environment to pursue personal freedom. Reflective autonomy is the foundation of innovation, as it empowers individuals to transcend conventional boundaries and shape unique, independent perspectives.

## **6. Transcendent Autonomy: the Moral Hero**

Transcendent autonomy presents a refined and highly coherent sense of identity operating in higher spheres (e.g., ethics) that align with universal truths. The integration of increasingly complex environments, in learning and metacognition, was modelled in our chapter 3 and demonstrated in chapter 5. This autonomy relies both on the larger cognitive perspective and on the ability for metacognition to refine and anchor of identity in these spheres. At this stage, the allocentric perspective evolves beyond binary distinctions which enabled to step away from an environmental construct now allows the individual to perceive themselves as a unit within a larger whole. At this level, metacognition refines identity's coherence in these higher spheres where action is driven while the value of the agent might be discounted. This perspective could present agents with an identity similar to a pawn on a chess game, a defined vector for change in a greater scheme.

The archetype of the Moral Hero exemplifies this level of autonomy, demonstrating exceptional coherence in intentions even when faced with intuitive or convenient alternatives, all for the greater good. Figures like Gandhi or Kepton Burton (*The Duke*) embody this autonomy as they display a profound sense responsibility when addressing societal or systemic injustices, prioritizing compassion and moral courage in the face of adversity.

## **7. Collective Autonomy: the Leader**

Collective autonomy relies on a higher sensitivity and conscious awareness of the gap between a high self-coherence and environmental coherence. This fine conscious access to the path for change provides agents not only with the ability to take voluntary action but also to communicate their motives to others. The highly refined coherence of the agent's identity (through metacognitive enhancement) may also provide the agent with a well-calibrated sense of overconfidence that is useful to navigate social environments.

While this conscious awareness might not necessarily build upon the allocentric perspective of moral heroes, we restrict our definition to the combination of both. Collective autonomy does not simply manage a team or teach knowledge but rather provides a vision for change in the higher spheres that can transcend others' identity

to differentiate, creating a vector for global sustainability and harmony on an entirely new magnitude. For this to occur, other agents must possess sufficient psychological development and resources to engage on equal footing.

Leaders are the archetype of collective autonomy. They not only create new value but also inspire others to adopt these motives as their own. J.F. Kennedy's leadership during the space race is a prime example: his ability to articulate a challenging vision "We do not do it because it is easy, we do it because it is hard" galvanized collective action and innovation despite adversity. Collective autonomy is essential for addressing global challenges, combining visionary thinking with the ability to mobilize others for collective change.

### **8. Extended Autonomy: the Alchemist.**

Extended autonomy envisions surpassing the biological limits of cognition by integrating agents with technologies (such as artificial intelligence or advanced science) that overcome bounded rationality. These advanced technologies may emerge from the collective efforts driven by the previous level of autonomy. Building on Chapter 6, metacognition in such agents could move beyond its thermostat-like function (which adjusts efforts to a desired level of coherence) and aim for optimal systems where perfect coherence could become possible.

The Alchemist is the archetype of such agents with a "supranatural" ability to both develop and transcend one's identity coherence into a legacy. Dating to the dawn of civilisation, alchemy sought a pseudo-scientific formula could transcend human limitations, such as mortality, by accessing higher levels of awareness and creativity. In such futuristic a scenario, human or cyborg decisions could align with more optimal, unbiased universal outcomes, potentially unlocking unprecedented coherence across systems.

We suggest that evolution has provided humans with, beyond the autonomy to navigate in social norms, the biological opportunity to act as vectors of change in ever more complex environments. Nested in the prefrontal cortex, this opportunity can be harnessed by the development of cognitive and metacognitive skills. Such autonomy with refined sensitivity to environmental and identity coherence could

foster voluntary action and challenge the contemporary issues we are facing in the real world. Indeed, while the UN's 17 sustainability goals for 2030 highlight the urgent need for humanity to develop unforeseen levels of autonomy through ingenious and goals-directed action. While this thesis began with the example of a modern Robin Hood relying on his explicit metacognition, we hope that the foundations laid here inspire further exploration of on how subjective value influences procedural metacognition to elevate humanity's potential.

#### **4. General conclusion**

We opened the thesis by asking what abilities heroes might possess to remain coherent (sometimes at great cost) with their values and goals. This interdisciplinary work provided a conceptual framework defining metacognition as inherently procedural and grounded in bounded rationality: a thermostat that ensures the reliability of decisions in light of the agent's goals. We presented a novel computational model and a study (in task with limited knowledge) to argue that monitoring signals ubiquitously track decision reliability and subjective value. Furthermore, we demonstrated that even in socially relevant tasks, such as moral decision making, metacognition is sensitive to decision coherence, as previously shown in hedonic tasks. Finally, preliminary results suggest that the metacognitive criterion of sensitivity to choice optimality might be a subjective trait across value domains, supporting the idea that this function must be accounted in a bounded rationality framework.

All together, we suggest that metacognition contributes to equipping humans with the ability to be autonomous and responsible citizen. It may also support the development of a sense of identity by enabling individuals to endorse a social role for the benefit of the group. We hope that this interdisciplinary thesis helps reframe research on metacognition to focus on its procedural aspects, which we believe have the potential to empower individuals in navigating complex environments and achieving their fullest potential.

## References

- Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and Control of Thinking and Reasoning. *Trends in Cognitive Sciences*, *21*(8), 607–617. <https://doi.org/10.1016/j.tics.2017.05.004>
- Bigenwald, A., & Chambon, V. (2019). Criminal responsibility and neuroscience: No revolution yet. *Frontiers in Psychology*, *10*(JUN), 1–19. <https://doi.org/10.3389/fpsyg.2019.01406>
- De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. *Nature Neuroscience*, *16*(1), 105–110. <https://doi.org/10.1038/nn.3279>
- Fleming, S. M. (2023). Metacognition and confidence: a review and synthesis. *Annual Review of Psychology*.
- Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2017). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), 17–19. <https://doi.org/10.1038/s41562-016-0002>
- 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), 54–61. <https://doi.org/10.1016/j.cognition.2010.06.010>
- Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
- Kitcher, P. (2011). *The Ethical Project*. Harvard University Press.
- Koriat, A., & Adiv, S. (2015). The self-consistency theory of subjective confidence. *Oxford Handbook of Metamemory*, *1*(June), 1–25. <https://doi.org/10.1093/oxfordhb/9780199336746.013.18>
- Lieder, F., Shenhav, A., Musslick, S., & Griffiths, T. L. (2018). Rational metareasoning and the plasticity of cognitive control. *PLoS Computational Biology*, *14*(4), 1–27. <https://doi.org/10.1371/journal.pcbi.1006043>
- Luhmann, N. (1977). Differentiation of Society. *Canadian Journal of Sociology*, *2*(1), 29–53.
- Maoz, U., Yaffe, G., Koch, C., & Mudrik, L. (2019). Neural precursors of decisions that matter—an ERP study of deliberate and arbitrary choice. *ELife*, *8*, 1–23. <https://doi.org/10.7554/eLife.39787>
- Moll, J., Krueger, F., Zahn, R., Pardini, M., De Oliveira-Souza, R., & Grafman, J. (2006). Human fronto-mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(42), 15623–15628. <https://doi.org/10.1073/pnas.0604475103>
- Pasquali, A., Timmermans, B., & Cleeremans, A. (2010). Know thyself: Metacognitive networks and measures of consciousness. *Cognition*, *117*(2), 182–190. <https://doi.org/10.1016/j.cognition.2010.08.010>

- Rahnev, D., Balsdon, T., Charles, L., Gardelle, V. De, Denison, R., Desender, K., Faivre, N., Filevich, E., Jehee, J., Rahnev, D., Balsdon, T., Charles, L., Gardelle, V. De, & Denison, R. (2021). *Consensus goals for the field of visual metacognition*.
- Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review, 45*(May), 31–51. <https://doi.org/10.1016/j.dr.2017.04.001>
- Rouault, M., Dayan, P., & Fleming, S. M. (2019). Forming global estimates of self-performance from local confidence. *Nature Communications, 10*(1), 1–11. <https://doi.org/10.1038/s41467-019-09075-3>
- Rouault, M., Seow, T., Gillan, C. M., & Fleming, S. M. (2018). Psychiatric Symptom Dimensions Are Associated With Dissociable Shifts in Metacognition but Not Task Performance. *Biological Psychiatry, 1–9*. <https://doi.org/10.1016/j.biopsych.2017.12.017>
- Schulz, L., Fleming, S. M., Dayan, P., Schulz, L., & Fleming, S. M. (2021). Metacognitive Computations for Information Search : Confidence in Control. *BioRxiv, 1–35*. <https://doi.org/10.1101/2021.03.01.433342>
- Seow, T. X. F., Rouault, M., Gillan, C. M., & Fleming, S. M. (2021). How Local and Global Metacognition Shape Mental Health. *Biological Psychiatry, 18*, 1–11. <https://doi.org/10.1016/j.biopsych.2021.05.013>
- Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences, 18*(4), 186–193. <https://doi.org/10.4324/9781315630502>
- Shea, N., & Frith, C. D. (2019). The Global Workspace Needs Metacognition. *Trends in Cognitive Sciences, 23*(7), 560–571. <https://doi.org/10.1016/j.tics.2019.04.007>
- Soutschek, A., Moisa, M., Ruff, C. C., & Tobler, P. N. (2021). Frontopolar theta oscillations link metacognition with prospective decision making. *Nature Communications, 1–8*. <https://doi.org/10.1038/s41467-021-24197-3>
- Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: Consciousness as an unconscious re-description process. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1594), 1412–1423. <https://doi.org/10.1098/rstb.2011.0421>
- Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two Key Steps in the Evolution of Human Cooperation. *Current Anthropology, 53*(6), 673–692. <https://doi.org/10.1086/668207>
- Vaghi, M. M., Luyckx, F., Sule, A., Fineberg, N. A., Robbins, T. W., & De Martino, B. (2017). Compulsivity Reveals a Novel Dissociation between Action and Confidence. *Neuron, 96*(2), 348–354.e4. <https://doi.org/10.1016/j.neuron.2017.09.006>

# Annexe:

## Glossary for the interdisciplinary study of value-based metacognition.

We propose here a lexicon that aims to provide consistent meaning throughout this interdisciplinary thesis (relating to cognitive neuroscience, economics, philosophy and some concepts in psychology). These concepts are mainly put to use to illustrate the theoretical framework (composed mainly by chapters 2 and 3) but are present throughout the entirety of the thesis from start to end.

### 1. Attributing value

#### 1.1. Decision type

**Value-based decision (VD):** the agent samples from their uncertain prior belief about the context payoff structure to define the item to be targeted as rendering higher value. Preferential tasks use an array of familiar items (whose payoff structure has previously been learned) and combine them in independent and random pairs (non-stationary context) to study economic decision-making. Reinforcement learning (RL) presents a few unknown items to study how participants adapt their strategy to stationary contexts with autocorrelated trials in which they learn the payoff structures through choices and feedback.

**Perceptual decision (PD):** the agent samples an uncertain perceptual signal from the local state of the world to identify the target item among two alternatives. PD generally present independent trials to cancel out the influence of a prior belief on the sampling from the local state.

#### 1.2. Learning

**Global:** relating to a stationary context's causal structure which defines the local states. In a controlled laboratory setting, a causal structure can be represented with autocorrelated trials (e.g. 98% of the time, the salt is in the more-dots shaker) or can be neutral with independent (and random) trials (e.g. 50% of the time, the more-dots shaker is on the right side). In value-based tasks, payoff structures are the causal structures that define the outcome of each alternative item. In perceptual tasks, dispositional structures are the causal structures that define the state in which the alternative items present themselves. The uncertainty of a causal structure can be probabilistic as for a lottery.

**Local:** relating to the present state of the world. Cognitive processes relating to local states are best studied in tasks with independent and randomised trials to cancel out global features of stationary contexts. In the world a local state is finite but can provide an uncertain signal due to its strength or noise.

**Bayesian learning:** the agent samples or acts in the local world to infer its global causal structure as prior expectations. Feedback together with the prior expectations about a decision form the predicted error (correct or wrong decision) which updates the prior with a learning rate.

### 1.3. Hierarchical decision

**Hierarchical decision:** decision in which both the global payoff structure and the signal from the local state of the world are uncertain. A hierarchical decision combines a value decision to define the target item and a perceptual decision to identify it in the local state of the world.

**Goal:** defines the currency to be valued and maximised when making a decision (*e.g.* accuracy, points, coins, welfare, salt...).

**Policy or decision-rule:** based on one's global belief, it defines the cue(s) on which one is to base sampling in order to identify the target item to choose: *e.g.* the agents targets the more-dots shaker because she believes it to yield the higher-value outcome most of the time, or she aims for the item on the right side because she believes that this is where the target item would generally be. Policies are learned as **prior beliefs** by sampling from or receiving feedback from a stationary context.

**Strategy:** value-based policy that defines the target item by associating one's goal with one's prior belief about the context's payoff structure (*e.g.* goal: value= salt; prior belief : salt = more dots ; strategy: more dots = target).

**Intention:** higher order preference that guides decisions.

**Decision rule:** resulting from inference (*e.g.* being explicitly given an instruction or learning from trial and error about the outcome of selected items and contexts), it defines the target cue (*i.e.* brightness, sugar, money, welfare...) on which one is to rely to select the correct option.

**Preference:** relative subjective value as revealed by decisions or verbal reports.

**Subjective value:** measure of how much an option satisfies the decision rule.



## 2. Decision and behaviour

### 2.1. Processes

**Reasoning:** explicit access to the intention guiding a decision.

**Dual process theory:** a framework in psychology dividing cognitive processes into two categories as either unconscious, fast, efficient and relying on heuristic or conscious, slow, effortful and reflexive providing the room for deliberate modulation of thoughts or actions enabling to overcome heuristic bias and therefore often result in a more rational behaviour. This definition of decision making systems based on a cost-benefit trade-off is conceptually analogous to our executive control dimension in the metacognitive landscape.

### 2.2. Decision quality

**Correct decision:** quality of a decision that maximises the decision rule either by coherently following subjective preference or accurately discriminating perceptual stimuli.

**Accurate decision:** quality of a correct perceptual decision, where objective features of the option maximise the decision rule, relative to option set.

**Coherent decision:** quality of a correct value-based decision where the selected option maximises the decision rule, relative to the option set. This decision can also be called optimal and relies on the correct perceptual discrimination of the present options.

**Satisfiable decision:** a decision where the selected option is not optimal but whose value is sufficiently high (and close to optimal) to satisfy the agent whose rationality is bounded.

**Decision reliability:** relating to the decision likelihood to be correct, accounting for objective features (e.g. variance or noise in the options' value or appearance) and/or subjective features (e.g. quality of the decision process as for response time or attention).

### 2.3. Behaviours

**Rational behaviour:** behaviour coherently maximising the utility of one's decisions in regards to her intention while accounting for uncertainty (e.g. probability, risk, volatility..)

**Authentic behaviour:** coherent behaviour that concerns a set of value relevant individual narrative and relying on self-awareness (e.g. moral preferences might qualify, colour preferences not so much: there might be several levels of authenticity like rings from core narrative to peripheral values note that we can hypothesise that the stakes are not the same with core authentic values as peripheral or non-core ).

**Autonomous behaviour: behaviour that remains coherent in context where there are counter intuitive options that differ from the coherent one.** It therefore relies on the S1/S2 trade off governed by M? agent who can control its own coherence with itself given its intention, across time with consistent and with other through virtuous cooperation.

**Sophisticated behaviour:** behaviour where the agent accounts for an upcoming challenge to choose in coherence with his intention due to the presence of more intuitive options (e.g. easier or safer) and takes anticipatory measures to discard these tempting but incoherent options. The famous example of Ulysses tying himself to the mast illustrates this behaviour.

## 2.4. Agents

**Maverick:** agent who can remain coherent to his own intentions in spite of social norms. Illustrative examples of mavericks are, as mentioned in the thesis, the character Truman (The Trueman show, 1998), or Nietzsche's ubermench Zarathustra who tend to isolate himself from society in order to remain true to himself and eventually develop his creativity.

**Hero:** agent with an outstanding ability to remain coherent to her intentions in spite of more intuitive options (easier, safer, more popular...) with the conscious and explicit aim to create value for the greater good. An illustrative example of moral hero mentioned in the thesis is Kepton Burton (a "modern Robbin Hood" portrayed in the movie The Duke, 2020).

**Leader:** agent who can remain coherent to his own intentions in order to create new value and is able to convince others to adopt these novel intentions as their own. An example of leadership can be seen as J.F. Kennedy who in time of need rallied a nation to develop new technologies and overcome adversity. His speech of 1962 illustrates his ability to lead through adversity "We do not do it because it is easy, we do it because it is hard".

### 3. Monitoring

**Metacognition:** (level-two) cognition about (level-one) cognition.

#### 3.1. Models

**Normative model:** framework regarding metacognition as monitoring system tuned to discriminate correct from incorrect decisions.

**Metacognitive inefficiency:** phenomenon where the normative model of metacognition is suboptimal by poorly discriminating between correct and incorrect decisions. A metacognitive inefficiency can be systematic if a given condition of the decision set up or process repeatedly affects the metacognitive ability (to discriminate correct decisions) throughout an experiment or throughout the literature.

**Procedural metacognition:** set of processes monitoring and controlling mental and behavioural states for flexible tuning of performance.

**System 1 - system 2 metacognition:** amongst other classifications of the complexity of metacognitive monitoring, the present classification divides them as respectively implicit (*e.g.* unconscious error correction or epistemic feelings that can be present in non-adult humans) vs explicit (*e.g.* conscious confidence reports).

**Metacognitive landscape:** proposed theoretical framework to observe various metacognitive functions along three cognitive axes of executive control, representation and conscious access.

**Metacognitive thermostat:** concept based on bounded rationality presenting metacognition (in the entirety of its landscape) as a monitoring and controlling system (*i.e.* procedural) that adjusts a criterion for an amount of resources (*e.g.* cognitive effort, time, dual process theory...) to be spent in order to reach an intended degree of cognitive or behavioural reliability.

#### 3.2. Monitoring signals

**Precision:** belief uncertainty that tunes its updating according to Bayes theorem.

**Prediction error:** binary expectation about the outcome of a decision as successful or not in reinforcement learning

**Epistemic feeling:** propositional states that suggest to an agent the likelihood of success of her decisions and can guide behavioural adaptation without requiring conceptual language.

**Confidence:** explicit report about one's subjective appraisal of a decision as correct (i.e. as either accurate or best given a set of options)

**Local confidence:** metacognition evaluates the reliability of a decision by sampling from its sources of evidence independently from the decision process. We argue that participants can distinguish two composite confidence levels in a hierarchical decision by monitoring independently the reliability of the chosen strategy (*i.e.* decision expected value sampled from learned prior) and of the perceptual identification (*i.e.* decision expected accuracy sampled from local perceptual evidence). Local confidence levels are best studied controlled laboratory setting with independent trials to cancel out global influences of a stationary context (*i.e.* generally perceptual and preferential decision-making tasks). Confidence levels are conventionally explicitly reported by human decision makers before an eventual feedback would be provided. In a psychometric model with stable difficulty, confidence computation can be calculated from the posterior probability of a gaussian representation as the evidence beyond the decision criterion supporting the choice (e.g. difference in dot number or value).

**Global confidence:** metacognition evaluates the reliability of a stationary context by sampling from its various sources of input (e.g. reliability of reward or of perceptual evidence).

**Retrospective monitoring:** monitoring signal appraising the reliability of a decision that has been made. These signals are known to account for features of the decision making process such as the response time. The monitoring signals are known to predict the repeatability of the made decision in the future.

**Prospective monitoring:** monitoring signal appraising the reliability of a decision that has not yet been made. These signals (either implicit or explicit) are suggested to be part of the process of reflection (alone or in groups) where detected low reliability can predict the investment of more time or resources (e.g. to sample evidence) before the decision is made.

**Declarative metacognitive knowledge:** accumulated knowledge about the reliability of one's decisions in a given context, depending on one's expertise or the context reliability. These appraisals are believed to be built thanks to feedback and education through childhood up to adulthood and to evolve throughout the lifetime of an agent.

# Curriculum Vitae

## Oriane Armand

### *Education*

- Since 2017    **PhD student in Systemic Neuroscience: Ludwig Maximilian Universität**  
Metacognition of value-based choices, Prof. Dr. O. Deroy.
- 2016            **MSc. in Advanced Neuroimaging: University College London**
- 2014-2015    Paris-Saclay University: 1<sup>st</sup> year of MSc. in Neuroscience.
- 2014            **BSc. Degree in Biology: Health option, Paris-Saclay University**
- 2013-2014    Imperial College London: 3<sup>rd</sup> year of BSc. in Biology, exchange student.
- 2012-2013    Paris-Saclay University: 2<sup>nd</sup> year of BSc. in Biology.
- 2011-2012    Lycée François I: 1<sup>st</sup> year of Prepa school of Biology (BCPST).

### *Publications*

J. Navajas, **O. Armand**, R. Moran , B. Bahrami, O. Deroy (2022) Diversity of opinions promotes herding in uncertain crowds, *Royal Society Open Science*.

**O. Armand**, B. de Martino, O. Deroy (in preparation) Confidence monitors and informs moral choices.

**O. Armand**, O. Deroy (in preparation) Inferential metacognition of value-based and perceptual decisions.

## *Awards and Fundings*

- 2015        **Erasmus + scholarship** for international summer placement with the University of Millan.
- 2013        **Erasmus scholarship** for full year international exchange program with Imperial College London.

## *Invited talks*

- 24/06/2021 **Inferential metacognition of perceptual and value-based decisions.** Metacognition new developments and challenges Conference, Zoom.
- 04/06/2020 **Confidence satisfices and guides choices in different value domains.** Berlin-Munich Seminar of behavioural economics, Zoom.
- 08/10/2019 **Confidence in preferences.** Graduate School of Systemic Neurosciences Orientation week, Martinsried, Germany.
- 10/05/2019 **Comparing metacognition across two value domains.** Agency, Consciousness and Metacognition Workshop, Barcelona, Spain.
- 21/07/2018 **Metacognition and motor action.** Graduate School of Systemic Neurosciences Retreat, Herrsching am Ammersee, Germany.

## *Invited Posters*

- 09/06/2021 **Confidence monitors and informs moral choices.** Life Improvement science online conference, MPI for Intelligent Systems, Tübingen, Germany, Zoom.
- 04/07/2019 **Charitable confidence: Moral options boost confidence in our preferences.** Summer School in Social Cognition, Aegina, Greece.
- 08/09/2018 **Does confidence track choice coherence across value domains?** Graduate School of Systemic Neurosciences Orientation week, Martinsried, Germany.
- 29/06/2018 **Estimates of cultural value that are believed to be shared with other are expressed with greater confidence.** Summer School in Social Cognition, Aegina, Greece.
- 28/06/2017 **Confidence: correctness or popularity?** Human Mind Conference, Cambridge university, United Kingdom.

## *Public engagement*

24-5/03/2018 **Public experiment: How do you think about art?** Tate Exchange program, Tate Modern London, United Kingdom.

26-9/04/2017 **Public experiment: Price of art and the value of discussion.** Tate Exchange program, Tate Modern London, United Kingdom.

## *Teaching*

03-6/2020 **PhD research projects tutoring: Research methods and data analysis.** Ludwig Maximilian Universität Munich, Germany.

03-6/2019 **Reading group coordinator: Computational models of decision making.** Ludwig Maximilian Universität Munich, Germany.

03-6/2018 **Co-teaching: Action theory in action: Introduction to experimental research.** Chair of Philosophy of Mind, Ludwig Maximilian Universität Munich, Germany.

03-6/2018 **Co-teaching and projects supervision: Research methods for philosophers.** Chair of Philosophy of Mind, Ludwig Maximilian Universität Munich, Germany.

10/17-02/18 **Bachelor project supervision: A new bank of value-based stimuli.** Ludwig Maximilian Universität Munich, Germany.

# List of Publications

J. Navajas, **O. Armand**, R. Moran, B. Bahrami, O. Deroy (2022) Diversity of opinions promotes herding in uncertain crowds, *Royal Society Open Science*.



