Graduate School of
Systemic Neurosciences
LMU Munich

The Challenges of Investigating the Sense of Agency by Explicit and Implicit Methods

Ondřej Havlíček

Dissertation der Graduate School of Systemic Neurosciences
der Ludwig-Maximilians-Universität München

SIGILL. UNIVERSITAT. LUDOVIC. MAXIM.

Munich, 13th of July 2016

Supervisor	Prof. Dr. Agnieszka Wykowska
2nd reviewer	Prof. Dr. Stephan Sellmaier
3rd reviewer	Dr. hab. Michał Wierzchoń
Date of defense	22nd of November 2016

Acknowledgments

I am thankful to a great number of people who made this thesis and my whole PhD project possible, friends, colleagues, and family, who contributed to my work directly or indirectly, spent long hours in discussions with me, gave invaluable advices, supported me in so many ways, were here for me when I needed it.

In the first place, I am deeply grateful to my supervisor Agnieszka Wykowska, not only for her supervision and introduction into the world of academia, but also for her never-ending enthusiasm, support, and friendship. My sincere gratitude goes also to Hermann Müller and Stephan Sellmaier for their supervision and countless advices.

I want to thank to all the other people who contributed to this work directly or through helpful discussions: Simone Schütz-Bosbach, Marcel Brass, Axel Cleeremans, Karl Friston, Antje Gentsch, Michael von Grundherr, Peter Hacker, Jakob Hohwy, Juraj Hvorecký, Simone Kühn, Floris de Lange, Tomáš Marvan, Dragan Rangelov, Davide Rigoni, Florian Waszak, and many, many others.

I also want to thank my colleagues but most of all friends, from the experimental psychology and neuro-philosophy groups, from the GSN and beyond, who are too many to fit on this page. Let me name at least Jairo, Joachim, Julia, Kelsey and Steffen.

I kindly thank the whole Graduate School of Systemic Neurosciences for the great and unique opportunity and stimulating environment and the financial support.

My whole PhD wouldn’t be possible without the love, help, patience, and encouragement of my parents and my wife Zuzana. Thank you!

Abstract

Since the beginning of the millennium there has been an increase of interest in the topic of the so-called sense of agency across many disciplines of cognitive sciences, ranging from philosophy, to psychology, neuroscience, or computational modelling. The term "sense of agency" stands for a variety of diverse phenomena connected with us being agents, having a body and performing goal-directed actions with consequences in the world and being aware of all of that. The research is largely motivated by pathological cases of impaired experiencing of agency, e.g. in psychotic delusions of control, alien and anarchic hand syndromes, or obsessive-compulsive disorder. However most of the research is being done with healthy participants.

It is generally acknowledged that the empirical research suffers from a lack of conceptual clarity and rigor, despite a lot of philosophical work that has been done on this matter. Many distinct phenomena are grouped under the same terms and vice versa. A related but under-appreciated issue concerns the methodology that is being used. That is the focus of the present thesis.

What we actually investigate depends on the methods that we use. While explicit methods use subjective reports, implicit methods aim to use objective data to infer the subjective experiences. In a theoretical chapter 2 *Investigating the Sense of Agency* I review selected prominent studies, showing that they often inform us about something different than what the authors think they do. Namely, in many situations, most of all when asking about unusual questions with no clear basis for answer, participants will translate the question as being about something more readily available, committing an attribute substitution. Many studies claiming to investigate subjective experiences in fact study various abilities which do not require any experiencing. There are however ways in which explicit reports can be very informative about the phenomenal experiences. Implicit methods, in turn, rely on many assumptions which I show to be difficult to meet. We can, after doing careful experimental work, infer the nature of the processes giving rise to an implicit measure, but inferences about subjective experiences are largely unwarranted, contrary to common claims. This applies to the popular sensorimotor implicit measures, sensory attenuation and intentional binding and in part to possible new measures, e.g. based on neuroimaging.

In the study *Metacognition of determinants of behavior: Learning to know more that we can tell* (chapter 3), we explored the reliability of explicit reports about the reasons why people performed a given action and of some other ratings. Many participants could not often tell automatic from deliberated actions. We provided half of the participants with a metacognitive training, during

which they received feedback messages about their decision judgments based on their response times. The training increased the metacognitive sensitivity of the trained participants but did not influence their response bias or other aspects of performance. Interestingly, most participants exhibiting rather good meta-metacognition: they knew well how much they can trust their metacognition. Moreover, we found that ratings of feeling of control over actions were strongly related to the inverse of ratings of difficulty across conditions, regardless of the actual control that participants possessed, which we interpret as attribute substitution. In sum, reliability of explicit reports depends on what the report is about and on the person, as there were large individual differences.

In the study *Expect to be distracted: Prediction of salient distractor by action and cue attenuates its interference* (chapter 4) we investigated the influence of action-effect prediction on attentional processing of salient but task irrelevant distracting stimulus as a potential cue and implicit measure for self-attribution of sensory events, similar to sensory attenuation. The intuition is that you do not need to pay attention to predictable, irrelevant, and potentially distracting side effects of your actions. Using the methodological framework of Hughes and colleagues (2013) we found that attenuation of the influence of distractor was probably driven by a more general predictive process and not necessarily by a self-specific action-effect prediction mechanism, such as the comparator model.

In conclusion, investigating the sense of agency by explicit and implicit methods is challenging, as is the whole science of consciousness. Plurality of approaches and interdisciplinary work are required. However, we do not need only more empirical results and theoretical and conceptual progress, but also methodological progress, to know how to answer which questions and what conclusions we can safely draw from our results. By properly employing various explicit and implicit methods, we can advance our understanding of the experiences, abilities, and neuro-cognitive processes connected with the notion of sense of agency. I hope that the present work can be beneficial in this respect to researchers studying not only the sense of agency but consciousness in general.

Table of Contents

1 Introduction

“The motion of our body follows upon the command of our will. Of this we are at every moment conscious.”

- David Hume, An Enquiry Concerning Human Understanding, 1748

“It is my hand and arm that move, and my fingers pick up the pen, but I don’t control them. What they do is nothing to do with me.”

- A patient suffering from delusions of control (Mellor, 1970)

1.1 The sense of agency

Few things in our lives are as commonplace as our ability to act according to our will. Philosophers have debated for centuries, with many cognitive scientists joining them recently, whether our will can be considered free and in what sense. While many authors can be skeptical about our free will, it is hard to question that most of us experience moments when we *feel like having* a sort of free will: We feel we have the ability to perform actions with a specific purpose in mind, to cause our body to move and to control this movement in order to accomplish a goal of our choosing. This and similar experiences are referred to as the sense or experience of agency. A more precise definition of this phenomenon will not be provided at this point, mostly because we are probably not dealing with a single experience but a multitude of various experiences, which are difficult to disentangle and capture under clearly defined concepts. They are related to various feelings and judgments about the kind of action that is occurring, in what way it is occurring (e.g. with physical or mental effort, without any attention to it...), who initiated the action, for what reason, who controls the unfolding or possibly inhibition of the action, who do the involved body parts belong to, and so on. Several philosophers and researchers have nevertheless undertaken such a conceptual investigation, which will be discussed later in this chapter. For now, let us consider what motivation there might be for occupying ourselves with these matters, why the sense of agency is worth investigating, both scientifically and philosophically.

1.2 Why it matters

It might seem as a matter of course that we perceive an action performed by our own limb as initiated by ourselves. We do not stop to entertain such strange ideas that it need not be so, that we could perceive the limb movement as guided by the will of the limb itself or that of our neighbor. One of the reasons for this is probably the fact that all the cognitive and neural processes that are involved work very well for most of us, most of the time, so we don't even notice or reflect upon them. However, one of the clearest demonstrations that these highly familiar and mundane phenomena are actually rather complex and need not work this way are certain psychiatric disorders: cases, in which some of the mechanisms "break down", leading to very peculiar experiences, some of which most of us probably cannot even imagine.

The most salient examples of disorders in the sense of agency can be found among the positive symptoms in schizophrenia. In delusions of control (also known as delusions of influence, delusions of agency, or passivity experiences), patients form false beliefs that some of their actions are not controlled by themselves and can attribute those actions to specific agents (like their relatives or neighbors) or external natural or supernatural forces (hypnosis, evil spirit). This perceived lack of control and external attribution can apply not only to actions, but to thoughts (delusions of thought control, thought insertion), feelings (delusions of emotional control, made feelings), and somatosensory experiences (somatic passivity) as well (Blaney, 2009). While these patients attribute events which can be caused internally to external influences, patients with delusions of reference (or, megalomania) tend to over-attribute external events to themselves (Synofzik, Vosgerau, & Voss, 2013).

Symptoms regarding misattributions of agency are not limited to schizophrenia patients. There are people who suffer from the so-called anarchic hand syndrome, in which their hand performs complex goal-directed actions, but this behavior is felt as involuntary and difficult to inhibit. The hand can be felt to have a will of its own, but is still felt as belonging to the individual. In contrast to that, a so-called alien hand syndrome can be found in people who claim that their hand in fact does not belong to them, and thus represents a case of partial hemisomatognosia (Marchetti & Sala, 1998). Moreover, patients with anosognosia for hemiplegia experience performing actions with their limb, although the limb does not move and they are in fact unable to move it (Fotopoulou et al., 2008). While these patients report awareness of an actual or imagined action, the patient GL suffering from complete haptic deafferentation, including loss of proprioception, "reported impressions of not controlling her movements, and not being aware what she was doing'', unless unbiased visual feedback was available to her (Farrer, Franck, Paillard, & Jeannerod, 2003, p. 616). There are other deficits on the border of awareness of some aspects of actions (the very occurrence of the action, reasons for the action) and ability to control actions, like utilization behavior, obsessive-compulsive disorder, Tourette syndrome, or narcotic addiction.

The diversity of all these disorders leads us to the conclusion that there is a complicated cognitive and neural “machinery” involved. Understanding this machinery will not only allow us to better treat these conditions, but also to learn very important facts about the normal experience of agency, in the sense of the ancient Greek appeal “know thyself”. Ultimately, understanding the sense of agency can be a crucial piece in our understanding of our very self-hood and self-awareness (Synofzik & Vosgerau, 2012).

1.3 The challenge

Regardless its theoretical and practical value, the investigation of the sense of agency may not be an easy task. Indeed, it might even be a “hard problem”, in the famous words of the philosopher David Chalmers (1995), because it is related to the problem of phenomenal consciousness. To many it seems impossible to have an objective science of first person experiences, while many others disagree that there is such a “hard problem” and suggest that investigating subjective experiences is entirely possible, although certainly difficult (e.g., M. A. Cohen & Dennett, 2011). The philosopher Daniel Dennett (2003, 2007) advocates the application of what he calls heterephenomenology, an approach he says scientists have been using for a long time: Collecting first person reports with all other available evidence and treating them like any other experimental data, subjecting them to careful analysis and interpretation, without taking the accuracy of the reports for granted. Obtaining reports from psychiatric patients suffering from delusions can be especially challenging. For this and other reasons, many scientists nowadays conduct research into the sense of agency with healthy participants. The hope here is to create laboratory conditions and settings that allow us to probe some aspects of normal experience of agency and thus learn something about the mechanisms that can be behind the abnormal experiences as well. For instance, it is possible to investigate under what conditions I can recognize observed hand motion as my own (Nielsen, 1963) or under what conditions I attribute some environmental effect to myself, as the cause of this effect, or in an opposite way, attribute effects of my actions to something or somebody else in the environment (Wegner, 2008). As much as these experiments are valuable, investigators need to be aware of the associated challenges, such that of the ecological validity of the studies, because many of the experimental settings are very unusual compared to everyday life and when confronted with such situations, people may have difficulties with providing reports accurately reflecting their experience. In connection to that, it has been recognized in the fields of cognitive and social psychology for some time that explicit reports about one's own psychological processes, such as what the reasons for an action were or how an action felt, can be very far from reliable (Nisbett & Wilson, 1977). To address this challenge, researchers have been searching for a substitute for these explicit measures, that is, implicit measures, which can be measured objectively, without relying on personal reports, but which in some way index one's subjective experience. It goes without saying that this task poses challenges of its own. For instance, it is difficult to ascertain what these implicit measures truly index and whether subjective experiences can be reliably inferred from them. And a very important issue, nowadays generally acknowledged but not always heeded, concerns the conceptual problems that are involved, as observed among others by another philosopher, Shaun Gallagher (2007). Many studies claim to investigate the "sense of agency", but what precisely are these studies capturing? We have already seen that this label groups highly diverse phenomena.

The general challenge thus is to deal with both methodological and conceptual problems in parallel to empirical research, if one hopes to arrive at a comprehensive theory of our experiences of agency. A more careful conceptual analysis of the investigated phenomena is needed in order to be able to interpret experimental data, build theories based on the data and also to design better future experiments. New experimental data will in turn provide input for the refinement of the concepts and methods used (Gallagher & Zahavi, 2008). A highly interdisciplinary approach will be needed, involving at least the fields of cognitive psychology, social psychology, psychiatry, neuropsychology, neurobiology, cognitive neuroscience, computational neuroscience, and philosophy. We now have a large body of empirical results and there have been important conceptual developments. In this thesis I want to argue that not enough attention has been devoted to methodological problems and that will be my main focus. I want to analyze some of the most commonly employed methods and see where their strengths and shortcomings lie. I will also report two empirical studies that try to take the methodological problems into account. But before we turn our attention to the methods and measures – how we want to investigate the phenomena – we have to briefly take a look at the theories and concepts: what it is we want to investigate.

1.4 Theories and concepts

In order to build a theory of any phenomenon – synthetizing experimental data into coherent explanatory and predictive frameworks – one must necessarily occupy oneself with not only experimental but conceptual work as well: carefully identifying the constitutive elements of the theory, their definitions, their relations with one another, and so on. By virtue of that, the resulting theory should not only be consistent with the empirical data, but also internally coherent, and should carve nature at the (most suitable) joints, as Plato famously wrote, i.e., not conflating phenomena that are better treated as distinct and unifying phenomena that are related. In this respect, the work of a scientist is very similar to that of a modern-day philosopher, although the scientist might not be aware of that. Moreover, philosophers are specifically trained in such skills as argumentation, conceptual analysis, maintaining a broader historical perspective, and with training in the cognitive sciences, they can greatly contribute to the scientific theories and conceptual frameworks (van Gelder, 1998).

Indeed, in the field of research on the sense of agency, there is a clear need for more rigorousness, conceptual clarification, and theoretical refinement, given the above demonstrated variety of possible deficits in the sense of agency and the variety of terms that are being used in the contemporary literature, see Table 1-1.

Table 1-1. Examples of terms and concepts associated with the sense of agency. *Partially based on (Pacherie, 2007).*

agency	awareness of a goal
self-agency	awareness of an intention to act
sense of agency	awareness of an urge to act
experience of agency	awareness of initiation of action
feeling of agency	awareness of movements
judgment of agency	sense of activity
sensation of agency	sense of mental effort
metacognition of agency	sense of physical effort
sense of ownership	sense of intentionality
experience of authorship	sense of initiation
experience of intentionality	sense of control
experience of purposiveness	sense of motor control
experience of freedom	sense of situational control

experience of mental causation	sense of rational control

These concepts are being employed to describe the experiences of participants in a wide variety of situations and tasks. However, without a clear conceptual framework, the concepts are likely to be employed inconsistently in scientific reports both between different labs and within the same lab in different studies.

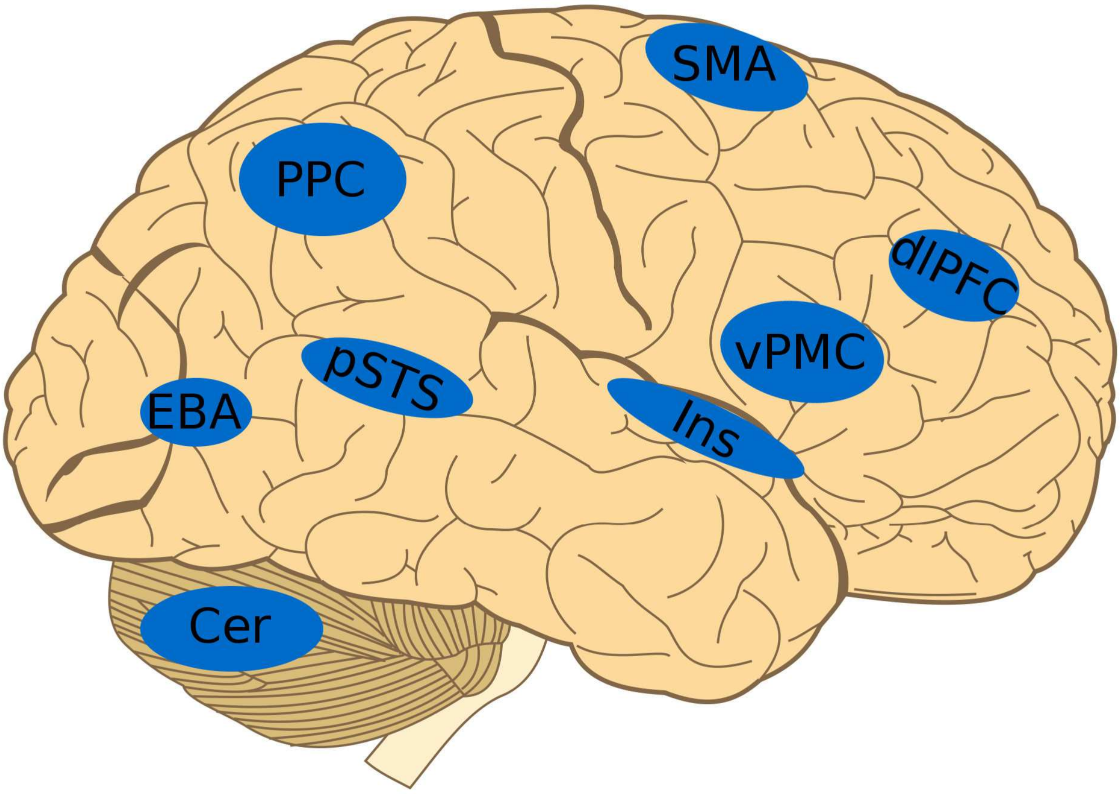

Figure 1-1. Brain regions implicated in the sense of agency. *These areas include the cerebellum (Cer), extrastriate body area (EBA), posterior parietal cortex (PPC), posterior superior temporal sulcus (pSTS), the insula (Ins), supplementary and pre-supplementary motor area (SMA, pre-SMA), ventral premotor cortex (vPMC), and dorsolateral prefrontal cortex (dlPFC). Adapted from (David, Newen, & Vogeley, 2008). This illustration of areas important for the sense of agency is likely to be incomplete, for instance because all sensory areas can provide agency cues and various (pre)frontal areas (e.g. the midline structures) together with temporal areas can be needed for high-level agency inferences. And importantly, it does not include information about the functional connectivity between these areas that are utilized in different aspects of the sense of agency.*

This can be seen not only in the search for the cognitive mechanisms but also the neural correlates of the sense of agency. Large portions of the whole brain have been implicated in “agency processing" (see Figure 1-1) based on rather different experimental manipulations, and systemizing these findings is a challenging task, which some have attempted (David et al., 2008, p. 530).

The systemizing the neuroscientific findings can be made much easier with a conceptual framework, an "ontology" of the phenomena in question, and therefore with a focus on the specific individual aspects and mechanisms of the phenomena. What can be seen as a first step in this direction is drawing a distinction between a sense of agency and a sense of ownership, as proposed by Gallagher (2000), similarly to Graham & Stephens (1994). Gallagher (p. 15) defines the *sense of agency* as "the sense that I am the one who is causing or generating an action", while the *sense of ownership* is "the sense that I am the one who is undergoing an experience." For instance, if someone else moves my hand, I have a sense of ownership for the hand, but not sense of agency for the movement. Gallagher (2007) discusses several neuroimaging experiments explicitly taking this distinction into account. He argued in a similar vein to Tsakiris and Haggard (2005) that all trials in those experiments should elicit the sense of agency, as no movements were involuntary, therefore there was no contrast specifically capturing that phenomenon, and more importantly, that the experimenters meant different things by the term "sense of agency" or "agency". While two of those experiments (Chaminade & Decety, 2002; Farrer & Frith, 2002) associate the sense of agency with intentional aspect of an action, i.e., pursuing a goal, another experiment (Farrer, Franck, Georgieff, et al., 2003) did not involve a goal-directed action, but focused on self-recognition of bodily movements and motor control. We can see this as an opportunity to develop a more fine-grained distinctions of the concept of the sense of agency (SoA): SoA as the experience linked to bodily movement and SoA as the experience linked to the intentional aspect of the movement (Gallagher, 2007). The conceptual work on this problem continues and is still far from complete (Gallagher, 2012; Pacherie, 2008; Synofzik, Vosgerau, & Newen, 2008). It also is noteworthy that the term "sense of agency" is often used on a higher, social level, for beliefs and facts that one's effort makes a difference in the world (Hitlin & Elder, 2007; Strahan, 2016). As far as I know, there is little interaction between the community using the term as referring to basic cognitive abilities and phenomenal experiences and the community using the term in this more social meaning.

As the title of this section points out, theories and concepts (distinctions, conceptual frameworks) are intimately related, if not inseparable. For this reason, not only does conceptual work lead to more refined theories (e.g., based on conceptual analysis or phenomenology; Gallagher & Zahavi, 2008), but theories of some related phenomena can help us pinpoint and associate some of the elements and processes involved in those theories with specific aspects of the phenomenology of agency. This contribution may be especially valuable from computational theories, as they are to some degree independent of higher-level psychological and phenomenal terminology, but can be tied to that higher level to some degree, as well as to the lower level of neuronal implementation, serving as an inter-theoretical bridge. This can be demonstrated by the application of motor control-based theories for explaining the sense of agency in health and disease.

1.4.1 Motor control-based theories

The historically most prominent group of theories relevant to the sense of agency deals with how our motor system works in relation to the perceptual systems. Hermann von Helmholtz (1867) has famously asked, when my retina registers a moving visual image, how does my brain "know" how to interpret this sensation? Does the external world move or is it only my eye that is moving? Similarly, why do two essentially identical eye movements lead to very different percepts: When I actively move my eye, I see the world as being stable, but if I gently push on the side of my eye bulb, so that they eye is moved passively, it looks like it is the world that is moving. What seems to be a crucial difference between an active and a passive movement is that there is a motor command in the case of the active movement, which can be used to figure out which sensations were caused by the self and which were not and even to alter our perception of otherwise identical sensory data. Through the work of such pioneers as Sperry (1950) or von Holst and Mittelstaedt (1950) we have arrived to a family of (optimal) motor control models (e.g., Blakemore, Wolpert, & Frith, 1998; Miall & Wolpert, 1996; Scott, 2004; Wolpert & Flanagan, 2001; Wolpert, Ghahramani, & Jordan, 1995; Wolpert & Kawato, 1998).

One possible model is depicted in Figure 1-2 (Synofzik et al., 2008). The agent forms an intention, a goal which is represented as the desired (perceptual) state of the world. A controller has to compute the best way (given some cost function or optimality principle) how to move the body in order to achieve this goal state. The controller is in essence a function from perception to movement, and is therefore often called an *inverse model*. The resulting motor command is used by the motor plant (e.g., an arm) to carry out the movement, which is also influenced by the environment, e.g. external disturbances. We then sample sensory feedback from the environment in order to estimate the resulting state of the world, including the state of our body. The estimated state is compared to the desired state and the computed motor error is used to improve the functioning of the controller and issue corrective motor commands (comparator 1). In parallel to that, a so called *efference copy* of the motor command is used as an input to another module, which tries to predict the resulting state of the world before it comes about. This prediction module is a function from movement to perception and is therefore often called a *forward model*. The predicted state, also known as *corollary discharge*, can be used for feed-forward control, before receiving actual sensory feedback (comparator 2). Finally, the prediction can be compared to the actual estimated end state (comparator 3) such that the predicted feedback can be attenuated and only the sensory discrepancy (*prediction error*) is perceived, such as when I perceive more strongly when someone else tries to tickle me, compared to my own attempt at tickling myself (Blakemore et al., 1998).

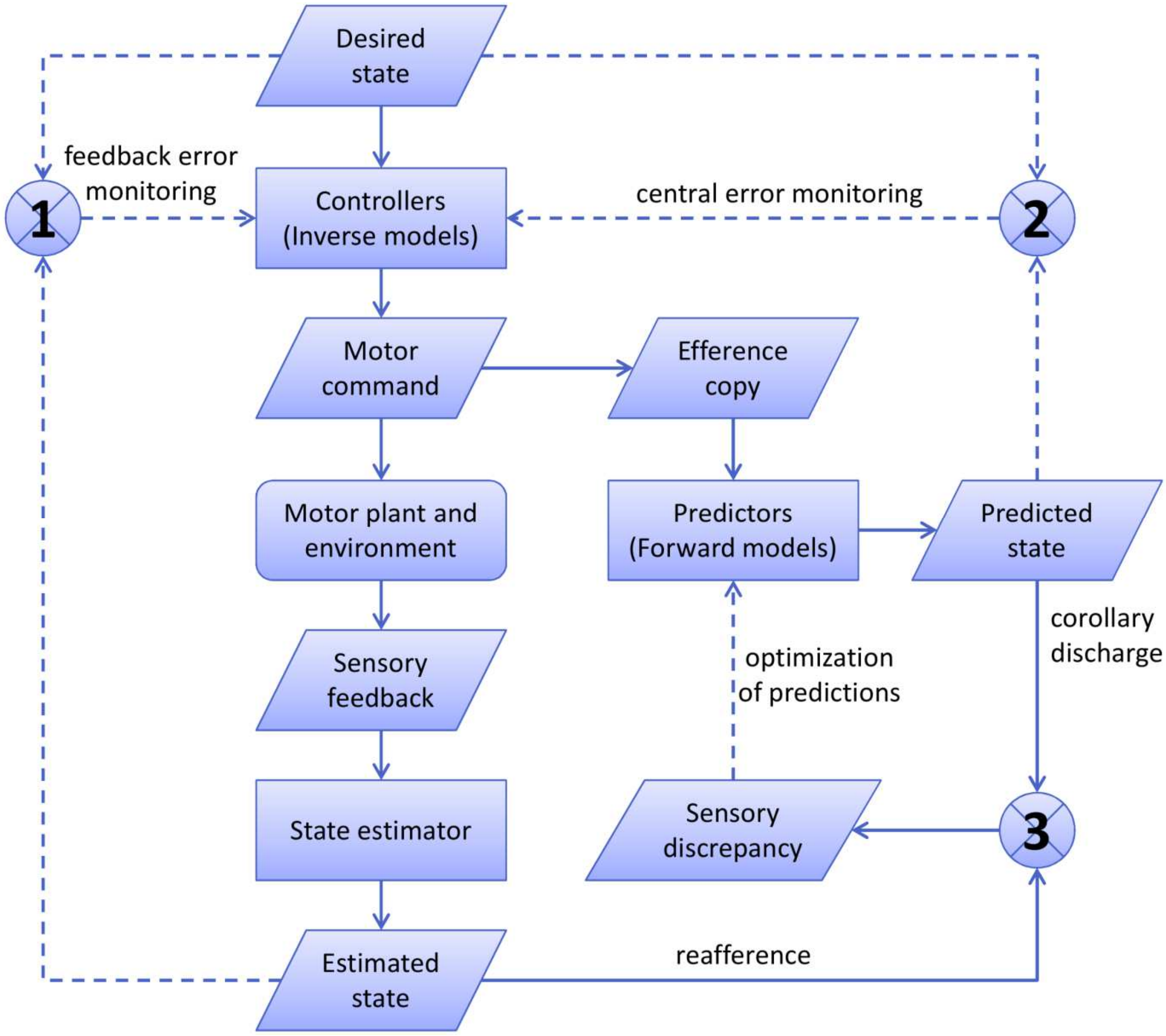

Figure 1-2. One possible model used in theories of motor control and the sense of agency. *Figure modified after Synofzik et al., 2008.*

The *comparator model* has been adopted beyond the domain of motor control, for explanations of deficits in the sense of agency (Frith, 1992, 2012; Frith, Blakemore, & Wolpert, 2000). According to this highly influential theory, some of our phenomenology of agency and associated deficits can be mapped to certain components and processes in this model. First, the processes involved in comparison of the desired and the predicted states (comparator 2) have been thought to underpin our *sense of control* over an action, as it allows online adjustments without the need for conscious control, and the action thus feels smooth and under control (Frith, 2005). And second, the comparison of predicted and actual sensory feedback (comparator 3) makes self-produced sensations feel differently and allows us to distinguish self- from other-produced sensory events, i.e., giving us a *sense of agency*. Note that we draw here conceptual distinctions (sense of agency vs. sense of control) based on the assumed underlying mechanisms. However, this model has been criticized as inadequate for explaining the most salient deficits in the sense of agency, such as delusions of control, in large part because of its narrow focus on low-level motor processes and disregard for higher-level cognitive processes (Synofzik et al., 2008).

1.4.2 Higher-level theories

In stark contrast to the low-level computational perspective, other theories approach the problem of the sense of agency, particularly self-attribution of events, from a higher-level, psychological perspective. In essence, the question of to whom I attribute an event is a matter of inference (conscious or nonconscious), which depends on my mental states, such as beliefs, desires, and intentions. While the comparator model stresses the aspect of prediction, these theories highlight (but are not limited to) the role of post-diction, inference, attribution, post-hoc rationalization, and the like.

For instance, Graham and Stephens (1994) explain the sense of agency for (self-attribution of) actions and even thoughts as depending on our intentional states, implicit theory of own psychology, and proclivity for self-referential narratives. “Thus, whether something is to count for me as my action depends upon whether I take myself to have beliefs and desires of the sort that would rationalize its occurrence in me.” (Graham & Stephens, 1994, p. 102) Emotional and motivational aspects can be strong factors as well. Sense of agency can be “shaped by affective appraisal of the actual action outcome” and “individual attributional styles” (Gentsch & Synofzik, 2014, p. 5). We have a bias for self-serving attributions, which means that “people tend to attribute positive events to their own personal characteristics but attribute negative events to factors beyond their control.” (Leary, 2007, p. 320) Moreover, it is often neglected (such as in the motor control-based theories) that the phenomena motivating large amount of research on the sense of agency – delusions of control in schizophrenia – often have “emotionally tuned semantic content” (Gentsch & Synofzik, 2014, p. 4).

Perhaps the most prominent higher-level theory of the experience of agency, “conscious will”, and agent-attribution of events comes from Daniel Wegner (Wegner, 2002, 2008; Wegner & Wheatley, 1999). Wegner goes as far as claiming that real causal efficacy of our thoughts is an illusion and that our experience of it is analogous to magical thinking: Just because there is a temporal relation between two events does not mean they are causally related (Wegner, 2008). Specifically, the first event would be a thought in our mind, such as an intention to do “X”, and the second event would be the occurrence of “X” in the world. According to his theory of *apparent mental causation*, whether we perceive ourselves as the causes behind an event depends on three conditions: (1) consistency: the thought and action must be consistent with each other, (2) priority: the thought occurs just prior to the action, and (3) exclusivity: there are no other plausible candidate causes. Wegner and his colleagues have shown that by manipulating experimental aspects related to these three principles, people can be led to self-attribute events that they did not cause and vice versa, to not self-attribute events they in fact did cause (Wegner, 2008).

However, people are sometimes unable to detect that the effect that had occurred was inconsistent with their prior thoughts (intentions), such as in the case of choice blindness (Johansson, Hall, Sikström, & Olsson, 2005) or real-time speech exchange (Lind, Hall, Breidegard, Balkenius, & Johansson, 2014), in which the outcome of the action (a choice of a picture or an verbal utterance, respectively) was replaced by the experimenters without the participants noticing it. These findings can serve as evidence against the comparator model as well, since it relies on the match of intention-based prediction and actual feedback. Lind and colleagues nevertheless argue in favor of higher-level inferential models in which multiple sources of evidence “are weighted in order to arrive at a conclusion whether the inserted word was self-produced or not” (2014, p. 6).

It seems likely that neither the low-level, nor the higher-level accounts are telling the whole story and that in reflection of the complexity of the phenomena of the sense of agency a more complex picture is needed.

1.4.3 Multiple cue integration theories

Matthis Synofzik and colleagues (2008) have proposed a multifactorial two-step account of the sense of agency involving both lower-level motor processes and higher-level aspects. They draw another conceptual distinction in the sense of agency between a *feeling of agency* as a “non-conceptual, low-level feeling of being the agent of an action” and a *judgment of agency* as a “conceptual, interpretative judgment of being an agent” (p. 222), similar to Gallagher’s (2007) distinction between first-order experiences and second-order reflective attributions. The feeling of agency (FoA) is supposed to only reflect whether an action is self-caused or not, without any specific external attribution, and the self is represented only implicitly. It results from weighting and integration of multiple indicators, such as internal prediction, sensory feedback and proprioception. It therefore draws on the motor control theories to a certain degree. If there is a congruency of the indicators, “we experience self-agency by a rather diffuse sense of a coherent, harmonious ongoing flow of action processing”. In the opposite case, “we experience an action as strange, peculiar and not fully done by me” (Synofzik et al., 2008, p. 228). The judgment of agency (JoA), on the other hand, is foremost a judgment, that means it is explicit, reflective, conceptual, inferential, and interpretative. When we form a judgment of agency, we try to come up with the best explanation as to who or what specifically caused the action or the event in the environment, that is, to rationalize the events. Importantly, this “rationalization does not depend on the comparator output and not even on reliable introspection, but rather on ad hoc theorizing about oneself”, on one’s intentional states, background beliefs, narrative self-structures, various social and contextual cues (Synofzik et al., 2008, p. 228).

There are therefore at least two levels of processing, involving a multitude of agency cues or factors. The difference in the cues can be also conceptualized in terms of the temporal relation to the action (Synofzik et al., 2013). There are sensorimotor and cognitive cues that are predictive of the action and its author, such as the forward model-based outcome prediction or higher-level anticipation. These predictive cues can influence the FoA, and through that also the JoA. There are also sensorimotor and cognitive post-hoc cues, such as the sensory feedback or the affective valence of the outcome that can influence our JoA, but also the FoA, and sometimes even “retrospectively” change our beliefs about what our prior intentions for the action were (Haggard, 2008; Kühn & Brass, 2009).

The fact that the complex phenomenology of agency should depend on multiple contributing factors can seem almost trivial. The question now arises, how should all the various cues be processed, what is their respective contribution to the multitude of agentive experiences and associated disorders? It has been suggested that the multifactorial weighting model is impossible to falsify, because any empirical results can be “explained” ad-hoc by some setting of weights, even zero weights for some factors (Carruthers, 2012). There have been suggestions that the weighting could depend on the relative reliability of each cue in a given situation, as in Bayesian cue integration (Moore & Fletcher, 2012; Moore, Wegner, & Haggard, 2009; Synofzik et al., 2013).

Indeed, Bayesian approaches to explaining various cognitive phenomena, including the sense of agency, are becoming increasingly popular.

1.4.4 Bayesian approaches

Recently a broad framework aspiring to explain a wide variety of cognitive and neural phenomena, even perhaps to be the grand unified theory of the brain and the mind, has been gaining prominence. It can be subsumed under the umbrella term of *predictive processing* (Clark, 2013) and is related to such notions as the *Bayesian brain hypothesis* (Friston, 2012), *predictive coding* (Rao & Ballard, 1999), and the *free energy principle* (Friston, 2010). I shall be using the terms "Bayesian framework" or "Bayesian approaches" for these notions collectively, because it is not clear yet which parts are essential, but the Bayesian aspect seems to be common to all of them, while the predictive aspect is present in the motor control-based theories as well.

Similar to the motor control theories, this framework can be also historically related to Helmholtz. In this picture, the brain can be thought of as an inference machine, which tries to provide us with a coherent experience of the world (perception) based on the various sensations, previous experience and current expectations. Bayesian statistics tell us how information from various sources (e.g., expectation, vision, touch, etc.) should be optimally integrated, according to their precision (or, equivalently, reliability or the inverse of the variance in the data) and it has been demonstrated experimentally that the brain can integrate information in a similar way (Ernst & Banks, 2002). The Bayesian brain hypothesis states that we have internal statistical models that try to represent the causal structure of the world. These probabilistic models can generate predictions, which are then tested on sensory data, which in turn leads to updated beliefs (empirical priors) about the causes of the sensations (Friston, 2012). Perception is essentially unconscious inference and the resulting percept is the best hypothesis accounting for the data. However, this is only a descriptive account and does not specify an algorithm or a neural implementation of the algorithm. It has been suggested that the computational principle of predictive coding can be one plausible implementation.

In a brief picture of predictive coding, there are multiple levels of computational units that try to predict the inputs from lower levels and it is only the *prediction errors* that get sent as input to higher levels (Rao & Ballard, 1999). According to Karl Friston (2005), the cortex is organized as such a hierarchical model, with lower levels representing low-level sensorimotor information (e.g., orientations, colors, etc.) and higher levels representing more and more abstract information (e.g., beliefs about abstract concepts, the world, myself, etc.). One important consequence of this picture is that there is no categorical or principal difference between percepts and beliefs, because both are just representations at different levels of the hierarchy. Another important aspect is the precision-weighting mechanism, which assigns different weights to prediction errors depending on several factors, such as the reliability of the input, our current goals (this can be conceptualized as "top-down attention") and ontogenetic and phylogenetic history ("bottom-up attention", but the distinctions makes little sense in this framework), see e.g. (Feldman & Friston, 2010). At this point it can be noted that the framework uses terminology from formal theories (e.g., precision weighting), but can be mapped to cognitive terminology to some degree (attention, confidence in sensory information), as well as to the neuroscientific terminology (modulation of postsynaptic gain of neural populations (Picard & Friston, 2014)).

Throughout life the hierarchical model should learn to represent the world we live in as accurately and simply as possible, minimizing its free energy, which is another important concept, the one which is according to Friston (2010) the basis for the unified theory of the brain. Under the free energy principle, organisms are thought to try to minimize the free energy of their internal model by perception, but also by action, such that they can sample inputs that can further minimize this quantity. It is in the domain of action, where the theory probably departs furthest from traditional views, that is, from the optimal motor control (Friston, 2011), by claiming that there is no inverse model, no controller that would perform complicated computations of the best motion trajectory etc. Instead, in a simplified picture, there is just one general model (Pickering & Clark, 2014), which causes movement by the principle of *active inference*, such that it changes precision weights for prioprioceptive and exteroceptive evidence and proprioceptively predicts that the body is already in the goal state, creating prediction errors, which are then resolved in a cascade via classical reflex arcs, resulting in the desired movement (H. Brown, Friston, & Bestmann, 2011).

With respect to the sense of agency, this model also combines the low-level aspects of sensorimotor prediction with higher-level aspects, leading to inferences about the author of the action (judgment of agency) on higher levels of the hierarchy and probably also to lower-level feeling of smoothly proceeding actions (feeling of agency) on lower levels of the hierarchy. While this explanation for the sense of agency in health is to some degree my speculation, Christopher Frith, the most prominent proponent of the comparator model, seems to be currently endorsing this framework with respect to explanations of various disturbances of the sense of agency (Adams, Stephan, Brown, Frith, & Friston, 2013; Fletcher & Frith, 2009; Frith & Friston, 2013). One of the strengths of the Bayesian approaches is the parsimonious explanation of positive symptoms in schizophrenia:

Hallucinations (false percepts) and delusions (false beliefs) are taken as fundamentally the same phenomena but occurring at different levels of the hierarchy. It has been proposed that there is a deficit in the precision-weighting mechanism (which is thought to rely on dopamine), such that prediction errors (sensory evidence) receive abnormally high weights, diminishing the role of priors. The sensory inputs are thus very hard to "explain away" at the lower levels, causing strange percepts, which results in the prediction errors propagating to very high levels, causing strange beliefs. Accommodating such persistent strange experiences can require changes in the whole conception of reality of the patient (Chadwick, 1993; Frith & Friston, 2013).

Furthermore, the framework can account for abnormal sense of ownership as well, in the case of the rubber hand illusion (Apps & Tsakiris, 2014). When we experience touch in our real hand, which is hidden from our sight, and at the same time observe the touch in the rubber hand in front of us, it can be a rational inference for the brain to (slightly and temporarily) change our body-model and assimilate the rubber hand. This is so because proprioception (which tells us that our hand is e.g. under the table) has lower precision than vision (which tells us that the perceived tactile stimulation is happening in front of us). Apps and Tsakiris even speculate that throughout our normal life simultaneously occurring multi-sensory sensations require a creation of an abstraction at a very high level, a bodily "self", to which these percepts belong and which can explain their harmony. I can speculate that by including action (and the resulting sensory feedback) into the picture, we can arrive at an abstraction of the self as an agent, with a sense of agency for own actions, at least in the sense of an ability to infer the likely agent (including self) behind an event, based on the various cues (see section 1.4.3).

However, it should be pointed out that the application of the Bayesian framework for explanations of the mind and brain is not without criticism, for instance because such models are hard to falsify (Bowers & Davis, 2012).

2 Investigating the Sense of Agency

We have seen why the topic of the sense of agency is an important one (section 1.2), we now also have a basic understanding of what it is we want to investigate and what theories we have so far (1.4). But what we actually end up investigating and learning about depends on the methods that we use. There are non-trivial challenges in this undertaking (1.3). Let us first divide the methods and measures roughly into two groups: explicit and implicit. Explicit measures consist of asking the participants explicitly about their experiences but also about objective facts, such as who produced some observed action. For that, we need to be sure that we are asking for the proper concepts, for example that we are not conflating the sense of agency with a sense of ownership. Additionally, it is clear that the responses will be judgments, therefore we cannot (easily, if at all) make inferences about the feeling of agency or other pre-reflective concepts. Implicit measures, on the other hand, consist of inferring the experiences from some objective measure that is usually associated with the concept we want to investigate, e.g. with experiences during a voluntary as compared to involuntary action. These measures can be derived from (predicted by) the theories of the mechanisms underlying the sense of agency, such as the comparator model, or can be discovered by other ways, but then are as well in a need of an underlying theory relating it to the investigated concept. It might be the case that the implicit measures are indicative of some pre-reflective experiences, like the feeling of agency. However, we will see that what the explicit and implicit measures actually inform us about can be something significantly different than what we think they do.

2.1 Explicit measures

Explicit measures seem as the most natural way of gaining knowledge about the sense of agency and related phenomena. If you want to know if someone feels as an agent, as the cause of some action, why don’t you ask her or him? There are many ways in which such inquiry can be done and consequently these will inform us about many different aspects of the phenomena. Consider several distinctions: We can ask people about their *internal subjective experiences*, their mental states, such as how fluent their action felt, how much they felt in control, how effortful it was, how responsible they felt for what they had done, what their intention behind the action was and so on. On the other hand, we can ask them about *external objective facts*, such as to discriminate (based on some available information) if it was them or someone else who caused some effect in the environment, if it is their or someone else's hand that they observe moving on a monitor (perhaps with some spatial and temporal distortion) and so on. It could be perhaps argued that the latter kind of reports investigate people's *abilities*, e.g. of self-identification, or causal inference, while the former reports investigate more directly the *quality of their experience*. (This difference could be roughly related to Ned Block's (1995, p. 227) controversial distinction between access consciousness: "availability for use in reasoning and rationally guiding speech and action", and phenomenal consciousness: "what it is like to be in that state".) While these distinctions may be useful methodologically, for making us think whether our research is aimed at investigating mainly the experiences or abilities of a person, the distinctions might not be always very sharp or meaningful ontologically, because an experience can be inseparable from some function (M. A. Cohen & Dennett, 2011), which can serve in the exercise of some ability. For instance, asking a participant if he or she was the agent producing some sensory event can be a question about both an internal experience (it is not publicly observable and may have a certain quality) and a fact (although sometimes there may not be any fact of the matter, as we will see later), revealing the quality of the experience, or an ability of the person, respectively. The internal/external distinction also need not be sharp, because even judgments about internal states can be formed based on external cues (Nisbett & Wilson, 1977). Things can get even more complicated when we recognize that some phenomena that we might consider as *experiences* are to a smaller or larger degree interpretative *judgments*, without much immediately felt phenomenal quality, such as when asking about how much someone felt responsible for some event. The word "felt", or "feeling" is leading us astray in this case, because the participant might not actually be performing any act of introspection of feelings, but an act of causal inference (how likely is it that I am the cause of the event?) or moral reasoning (is this an action for which people are generally held responsible in this society and what are my personal moral convictions?).

This leads me to a larger point I want to make: When people are asked about things that are not the typical things that they deal with in their everyday lives, about concepts and terms that do not form part of their ordinary language, about things that are too complex, or are even improper questions, ambiguous, not specified in behavioral terms, or when there is no fact of the matter, no correct answer, but one is required, I want to suggest that people are likely to try to understand the experimental situation in an idiosyncratic fashion, to make some sense of the question for themselves, translate and re-interpret it, even commit a so-called *attribute substitution error* (Kahneman, 2011; Kahneman & Frederick, 2002), which is a "heuristic in which a difficult question is answered by substituting an answer to an easier one" (Kahneman & Frederick, 2002, p.

50), even without being aware of that. I propose that this could present an important problem in the research on the sense of agency (and possibly other fields of study) and aim to demonstrate it via analysis of selected experimental literature and in my own experiment (see chapter 3).

2.1.1 Validity of agency judgments

A widely cited[1] study by Atsushi Sato and Asako Yasuda (2005), titled *Illusion of sense of self-agency: discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership*, aimed to "directly investigate whether the prediction of sensory consequence of actions made by forward model would modulate the sense of self-agency" (p. 243), defined and operationalized as "it is I who am producing the tone" (p. 251). Let us now put aside that the part "made by forward model" is unnecessary (the essence of the sentence and the research itself would be the same without that part) and somewhat speculative (the forward model is a theoretical entity and it is not certain if there is something sufficiently similar to that in the brain and whether it can predict *exteroceptive* sensory consequences of actions) and analyze the tasks, results, and interpretations in some detail.

In the first experiment of the study, participants first learned an association between two actions (press of a button with their left or right hand) and two effects (600 or 1000 Hz tones) following the actions immediately. In the test phase, participants were again pressing the two buttons and subsequently hearing tones, but the tones could be either the same (congruent) as the tone they learned normally follows the button press or the other tone (incongruent). Also, the tone could follow the action immediately, as participants learned it normally does, or after a delay (200, 400, or 600 ms). We learn from the description of the study that "participants were told that there were two cases: in one case they might hear tones as a result of their button press, but in another case the experimenter might produce the tones" (p. 224), but it is not clear how or why this should be so, i.e., there is no description of whether there was some additional person and a story of how that person's actions could override the action-effects of the participant. After each trial (button press and a tone), participants were asked to answer two questions on a scale from 0 ("totally disagree") to 100 ("totally agree"): "I was the one who produced the tone" (supposed to capture the "sense of self-agency") and "I was the one who was listening to the tone" (supposed to capture the "sense of self-ownership").

[1] As of the 14th July 2016, Google Scholar registers 227 citations.

The second question received not surprisingly ratings of 100 in all participants and conditions and let to the conclusion that “the predictability of the sensory consequences of actions ... does not always affect the sense of self-ownership” (p. 245). While the question is arguably curious and not very informative in relation to the conclusion (why should my ability to identify that it was my body parts that were pressing the buttons be influenced by what happens after the movement?), it is at least meaningful, there is a clear fact of a matter to it. With the first question (whether it was me who produced the tone) it is more complicated. The tone was always produced as a result of the actions of the participants, there was no experimenter pushing own buttons, playing in some tug of war or overriding the participant’s control over the tones. But we can equally well say that it was always the computer “who” produced all the tones. The computer “saw” that the participant performed an action, but it was it “who made a decision” which tone would be produced and when. Knowing this, what would be the *correct* answer to the experimenters’ question? Always 100, because I was always causally implicated, or always 0, because it was always the computer who made the final call about what would happen? Or to give no answer at all, revolting to the experimenters, because there is no clear correct answer? Of course, the participants want to cooperate and provide an answer. They can even believe that there is indeed a second person sometimes producing the tones. However, even if there were truly a hidden experimenter, sometimes pressing his or her “override” button, resulting in the silencing of the participant’s tone and producing his or her “own tone” instead, the only information available to the participants, based on which they can make the correct judgment about the author of the tone, is the pitch of the tone and the delay (not to mention that even that is not certain, because the experimenter’s tone could happen to be the same tone as which would be produced by the participant and at the same time, leaving absolutely no basis on which to make the judgment).

Crucially, at this point, the experiment is equivalent to a psychophysics experiment, in which people are required to perform two tasks: to discriminate two possible frequencies of tones and judge the length of temporal intervals. When people are then asked how much they agree that it was them who produced the tone, the only thing they can do with the available information is to (unconsciously or consciously) *translate the question as being about this available information* and additively combine the two psychophysical judgments and map them somehow onto the response scale. This is precisely what the results showed. Hearing the previously associated tone at zero delay resulted in on average “self-agency” rating near 100, while hearing a different (incongruent) tone subtracted ca. 15-20 points from this rating and each additional 200 ms in delay subtracted ca. 10-20 points (my estimates based on the depicted data; regression analysis was not reported).

The authors conclude from the results that “the predictability of the sensory consequences of actions modulate the sense of self-agency” (p. 245). I argue that because of the above mentioned problems such an experiment does not tell us much, if anything, about the sense of self-agency as a “modulable” experience and only a somewhat trivial fact about the sense of self-agency as an ability. With respect to this ability, I maintain that the experiment does not tell us anything about the truth or falsity or details of operation of the forward model as the means for action-effect self-attribution, as is claimed. Excluding alternative explanations for supposed effects of the forward model is difficult in general (Hughes, Desantis, & Waszak, 2013). And in this case the alternative explanation can simply be that the ability is to perform certain psychophysical discriminations and try to make some sense of the experimental situation, instructions, and available information by using the detected differences in the stimuli and various own assumptions how this should be mapped onto the response scale. (E.g., if some people assumed that any difference in the stimuli is a result of the other person’s agency, the “appropriate” agency rating for them could be 0, not a number corresponding to additively combined congruency and delay information. Some other people might respond always with 0, some always with 100, some always with 50, as discussed above. It would be interesting to investigate data of individual participants to see if different participants had these different types of mapping. It would also be interesting to know if some participants expressed uncertainty as to how to understand the task and to make responses.) As a possible test of my interpretation, I predict that we could obtain similar results in other conditions, where we would vary different information compared to the previously learned action-effects (such as shades of colors accompanying the tones, etc.), where we would provide much fewer learning trials (just enough to learn the “standard” stimulus for psychophysical discriminations), or where the participants would not act at all and were just observing another person (or just a cartoon robot/etc. on a screen) act and would need to judge whether it was this person/robot or someone else who produced the tone.

Second experiment in the study added an “actual agency” condition, such that sometimes the tones were the result of their own agency, sometimes of the computer’s. Participants pressed the two buttons in response to two color stimuli (i.e., a choice RT task) and heard again one of the two tones. In the “other agency” condition, the mean response time from a training period was used to estimate when the participant could normally press a button, and the tone was presented after some delay from that time point. In the “self agency” condition, this time point was the time of the button press as usual. Again, tone could be congruent or incongruent and the delay could be 0, 200, 400, or 600 ms. Importantly, only those trials in the “self agency” condition were included in the analysis, in which the reaction time was within 15 ms from the mean reaction time (which was the basis for the “other agency” condition). One can see that in this experiment the timing and the pitch of the tones were again always determined by the computer. More importantly, there was again no basis on which to make a factual agency judgment, because the information available to the participants was the same as in experiment 1. Moreover, from the empirical point of view, there was virtually no difference between the self and other agency conditions. In both conditions the tones were presented after the constant mean response time of the participant (± 15 ms) plus one of the possible delay durations. It is thus not surprising that the rating patterns in both agency conditions were almost exactly the same as well. Nevertheless, the authors interpreted the results as showing that participants misattributed their own actions to “the other” and also misattributed “the other’s” actions as their own, which could be compared to being surprised that I cannot recognize which of two identical coins was the one which had been previously blessed by a priest. This is a comparable case of essentialism, because the participants were supposed to recognize the essence of the computer algorithm that was behind identical sensory effects following in identical ways identical actions. The authors also used these results to draw theoretical conclusions that “the sense of self-agency was built on the comparison between the sensory prediction made by forward model and actual sensory feedback or on the comparison between intended and actual consequence” (p. 248). Again, I find these conclusions largely unsupported by the data and more importantly, investigated using unsuitable experimental paradigm.

Sato and Yasuda conducted a third experiment to investigate further upon what comparison (actual effect with the predicted or intended effect) the sense of agency is built. In this experiment, the tone was presented in relation to the participants’ response to a flanker task (Eriksen & Eriksen, 1974), that is, a discrimination task in which it is easy to make an error and respond with the other of two possible buttons. Participants were asked on each trial if they think they had made an error or not. The authors’ reasoning was that correct responses present cases of congruence between actual effect and intended effect, while incorrect responses identified as incorrect present cases of congruence between actual effect and effect predicted by the forward model based on a motor command. That is, any difference between these two classes would be judged as being due to a lack of intention to produce the heard tone in the case of an error, rather than just being due to the simple fact of making an error. This represents (another) serious conceptual problem of the study, making it impossible to draw any conclusions regarding the presented research questions, regardless of what results one finds.

The obtained pattern of results was similar to the previous two experiments. There was no statistically significant difference between the "intentional" (correct) and "unintentional" (error) response conditions, mainly because there was virtually no difference in self-agency ratings between the cases of "deviant" stimuli (incongruent tones and non-zero delay). The only difference between "intentional" and "unintentional" conditions was in the one case of "standard" stimulus (congruent, zero delay), where being correct contributed additional 20 "agency points", toward almost 100. Overall, the assumed intentionality or unintentionality of the response made remarkably little difference. However, the fact that the correctness ("intentionality") does not have an independent additive effect and seems to selectively add to the ratings for the correct rather than subtract from those for incorrect standard effects (tones) makes the interpretation quite difficult. The authors' interpretation was that: "These results suggest that the sense of self-agency depends on a comparison between intention and actual consequences of movements but does not totally depend on it. Rather, it seems that the sense of self-agency might mainly depend on a comparison between the predicted and actual consequences of actions." (p. 250). Again, a more plausible interpretation seems to be that the participants were again doing psychophysical discriminations, with additionally taking into account whether they had made an error, and mapping the information onto the agency scale using various own assumptions (see above my discussion of the first experiment). Individual differences in the assumptions and understanding of the experimental situation can also explain why the pattern of responses was the same in all three experiments but the actual values varied.

The paper raises other concerns such as about the authors' interpretations of the constant ratings of the sense of ownership, which I will not discuss in detail. The main moral we can take from this study is that *investigation of the sense of agency is difficult*, methodologically, theoretically, and conceptually. (The same goes for matters of consciousness in general.) It is a complex problem that requires knowledge in several subject areas and careful dealing with seemingly minor issues, which is arguably challenging in today's academia, with known pressure for publications (Brischoux & Angelier, 2015) and strong conclusions (Vinkers, Tijdink, & Otte, 2015).[2] The problem of complexity applies to the present author as well, who may very well be wrong in his analysis, here and elsewhere.

[2] Let my speculative idea be noted at least in the form of this thesis: I am suspecting that there is a possibility of a "research bubble" in at least some portion of the field of the sense of agency, similar to an economic bubble, driven in part by the popularity of the field. I by no means imply that this concerns all or even majority of the research. Additionally, perhaps my idea stems from lack of information and understanding, which is certainly possible, and in that case I apologize to the field and individual authors.

The study has become accepted and rather prominent in the field. I am not aware of any critical reflection on it. A review exclusively of this study (Knoblich & Sebanz, 2005) agreed that the “experiments show that intentions and motor predictions contribute to the experience of agency” (p. 261). Knoblich and Sebanz interpreted the self-agency ratings as being about the felt intensity of the experience of agency, although I would argue that the ratings capture an ability rather than an experience (as discussed above), showing that the same measure can be understood in quite different ways and that the notions of experience and ability I discussed above can be usefully distinguished. They nicely point out that it is surprising that the pattern of self-agency ratings did not differ much between the first experiment, in which the actions were self-chosen, and the second experiment, in which participants were instructed which button to press. Intuitively, we would expect that if the ratings are really about some aspect of what it is like to be an agent, the ratings should be much higher for freely chosen actions. The reviewers try to explain this surprising fact by pointing to lower ratings in the cases with temporal delays in the forced-action experiments, and there can be something to it, however, my explanation is that the participants did not attempt to assess the strength of their agentive experience but rather guess who the agent was and in the more complex situation of experiments 2 and 3 and for the specific participants it seemed more implausible that they were the agents if there was any noticeable delay. Knoblich and Sebanz also consider alternative explanations of the results, namely Wegner’s post-hoc evaluation account (see section 1.4.2) or error monitoring account, but still within Sato’s and Yasuda’s conceptualization of the experiments, such as that the manipulation in the experiment 3 is about intentions, or that the ratings are about experienced agency. This makes it difficult for the reviewers to explain some problematic points that they identify, such as how it is possible that there could be a high experience of agency even when the action effect is not consistent with the intention (“priority can be sufficient for the feeling of causing an action”) or for erroneous actions (“error-monitoring signal is used to readjust the system” and this “could serve as a direct indication of agency”, p. 260).

Other researchers have based their studies on the experimental paradigm of Sato and Yasuda. Simone Kühn and colleagues have conducted a study investigating the EEG/ERP correlates of processes leading to the authorship ratings (Kühn et al., 2011). They employed the same paradigm as in the first experiment by Sato and Yasuda (2005) with one of the small differences being usage of 100 ms delay instead of 0 ms for the “normal” action-effect association and 300 and 600 ms for the other delay conditions. The story for the possibility of the tone being produced by an experimenter was more credible, because the experimenter was seated close to the participant (although the experimenter was again never involved in production of the tones). The dependent measure was a rating for the question “Who produced the tone?” on a scale from 1 (“Me”) to 100 (“Somebody else”). Therefore the framing was slightly different from the original studies, where the question was posed in terms of agreement or disagreement with a statement that “I was the one who produced the tone”. The pattern of their results was again the same as in the original studies. However, the mean ratings in all conditions were remarkably close to 50 (interpretable as “I don’t know”), with the previously associated tones and delays receiving a mean rating of only ca. 42, 300ms-delay-tones receiving a rating still more favoring the “me” rating (around 46), 600ms-delay-tones receiving rating slightly favoring the “somebody else” rating (around 55), and the difference in ratings for congruent and incongruent tones being remarkably small (ca. 2 points). These differences from the original studies can be the evidence of how much the differences in framing of the question, in the individual participants, and perhaps in cultural context influence the supposed measure of the sense or judgment of agency.

Of substantial interest are the ERP results. In an important validation of the action-effect association, the N1 component (locked to the onset of the tone) was found to have lower amplitude when listening to congruent compared to incongruent tones, suggesting the occurrence of sensory attenuation for predicted action-effects (see sections 1.4.1 and 2.2.2), which really might be due to a forward model or some other prediction mechanism. The authors also analyzed ERPs for tones that received above-median versus below-median authorship rating, thus asking whether there are electrophysiological correlates that can predict the rating to some degree. They report analysis only for trials with congruent tones and 300 ms delay, because according to the authors these tones present the hardest case to detect mismatch from the learned action-effect association. For these “difficult” trials the ERPs revealed significantly higher amplitude of a P3a component for the trials which were rated above-median as produced by “somebody else”. The authors suggest that in the case of augmented P3a the tones were processed as “odd”, more unexpected compared to cases with smaller P3a amplitude, and that the 300 ms delay could have been perceived as being different or similar to the normal associated 100 ms delay depending on fluctuation of attention, mind-wandering. Kühn and colleagues interpret this result as showing that “agency judgments incorporate early information processing components within the range of the evoked potential and are not purely reconstructive post-hoc evaluations generated at the time of judgment” (p. 6). I am not sure if these two options are mutually exclusive, since post-hoc evaluations draw on many cues (Synofzik et al., 2008), likely including perceived differences in timing. Importantly, the fact that the tones perceived as different from the normal associated tones received higher ratings also fits with my psychophysical discrimination and attribute substitution interpretation, or in other words, that the ratings do not directly reflect perception or experience of being an agent.

An experiment by Farrer and colleagues (Farrer, Valentin, & Hupé, 2013), which had a different and larger aim that what I am discussing here, represents a similar case where participants needed to subjectively interpret what action-effect delays should be mapped to what rating responses. In this case there was only one action, the effect was a grey ball displayed on a screen, and the delay could range from 0 to 1100 ms in 14 steps. The action-effect delay was always decided by the computer, but participants were told that “either the ball would appear directly after they had pressed the button, or it would appear after a certain amount of delay, or the computer itself would control the ball and in that case they would have no control over it” (p. 1433). There were therefore only three possible response ratings (choices): “(1) Self: ‘‘my button press directly triggered the ball’’; (2) Delay: ‘‘my button press triggered the ball but it appeared with a time lag’’; (3) Other: ‘‘my button press did not trigger the ball, it’s the computer that triggered it.’’ (p. 1433), named as “full control”, “partial control”, and “no control”, respectively. The decision between choice one and two should arguably depend on the ability to detect a non-zero delay. However, there does not seem to be an objective basis on which to make the third choice, because it is not specified how long the lag after my action in the “partial control” case can be, and what is the lag distribution in the “no control” case. The participants thus have to make some assumptions about these matters, i.e., decide what delays should be mapped to which of the three ratings.

The design is more complex and the results likewise, reflecting the larger question of the study. However, it is interesting to observe that the mapping between the delays and the three rating choices can be modelled by three functions (with delay on the x-axis and proportion of selecting the given choice on the y-axis), one for each rating choice: A decreasing sigmoid function for the full control responses, hill-shaped function for the partial control responses, and an increasing sigmoid function for the no control responses, with inflexion-point boundaries at ca. 334 and 708 ms. It is interesting that these boundaries parcel the “delay space” roughly into equal thirds and that the slopes of the functions are relatively shallow (small), even for the largely objective discrimination problem behind the “full control” category, arguably reflecting an uncertainty as to what the appropriate response should be given the experimental situation and available information. It is certainly interesting, although not surprising under my interpretation, that the participants interpreted the situation such that the ‘computer itself controlled the ball’ much more often when the delays were in the last third of the delay space rather than in the second or the first third, even though according to the instructions the computer might have produced the tone after any delay.

The case for questionable validity of the agency judgments in such tasks can be made even stronger when we consider situations in which there is no action involved but an agency rating is nevertheless required. We can find such a situation in the experiments of a study by Sato (2009), which also had a larger scope that is not a matter of the current analysis. In this study there were conditions in which participants performed actions and conditions in which they merely observed the experimenter's actions (and tone effects), with various manipulations. Interestingly, when asked a question "To what degree did you feel that you were the one who produced the first tone?", on a scale from 1 ("not at all") to 8 ("very much"), participants provided ratings well above one (means between ca. 2 and 4) even for the cases in which they had not acted, which again raises a question about what such ratings actually reflect.

There are other classes of experiments in which such issues could be manifested. There is a multitude of studies in which participants performed some type of action and perceived some type of effect, where the properties of the effect (such as its perceptual features, timing, congruency with action or with another stimulus) were determined always by the computer, but participants were asked for some type of agency or control judgments (e.g. the degree to which they thought their actions determined the properties of the effect). While there are important differences among these experiments and each would require a separate analysis and discussion, a common feature is that there is often no fact of the matter to be rated[3] or how the rating should be provided is ambiguous or not well-defined[4], but people are still required to provide a rating. Therefore they will try to arrive at the rating using the available information, consciously or unconsciously, in uncertain ways (not reliably determinable by the experimenter and perhaps the participant as well). While the determinants and ways in which people arrive to such judgments can be certainly interesting, adequate interpretation of the results is more difficult than can seem, as I have demonstrated above. Specifically, it is hard to make claims about factual phenomenal experience, i.e., that participants *really* "felt" some "experience of agency" or "had" a "sense of agency" or of "control". What we usually can claim is that participants exhibited certain cognitive abilities and some systematic mappings between available information in the experiment and the responses, and the important

[3] E.g., asking for a "control" rating when the participants cannot influence (control) what happens in the task through their effort or asking for authorship ("agency") when there is no other agent, while the outcomes are determined purely by a computer in some fixed or random fashion. See the discussions above.

[4] E.g., asking for a "control" or "agency" rating, instead of asking e.g. in what percentage of trials a given effect follows a given action. Using such an "objective" framing of the task, the participants are not forced to perform a translation, an attribute substitution of sorts, between the "objective" fact (percentage of action-effect congruent trials) and the response scale provided by the experimenters. The results are then more easily interpretable as being about a particular ability, rather than a vague concept, probably not easily understandable by many participants.

question here is determining the nature of the abilities and of the mapping and what their theoretical significance is. As I have suggested, the answer can be interesting (e.g., that probably even subliminal information predicting the action-effect can influence the judgments; Gentsch & Schütz-Bosbach, 2011), often very hard or impossible to arrive at (to adequately interpret the results), and often can be rather trivial, such as when the only varied information is the delay between actions and effects (shorter delays will likely receive higher agency/control ratings) or when the effect more reliably follows (is more congruent with) the action (higher congruence will likely receive higher ratings), or when different conditions lead to different numbers of errors (condition with fewer errors will likely receive higher ratings).

2.1.2 Task objectivity and report factuality

There does not seem to be a clear boundary separating experiments in which the ratings are clearly interpretable as valid with respect to the research questions and those in which this is somewhat problematic, as shown above. For instance, compared to experiments in which the effects of actions are random to some degree, in another popular experimental paradigm (Valérian Chambon, Moore, & Haggard, 2015; Valerian Chambon, Wenke, Fleming, Prinz, & Haggard, 2013; Wenke, Fleming, & Haggard, 2010) the action effect (circle of a specific color) is objectively determined by the participant's response (left or right key) and what precedes this response (a cue telling the participant which response to choose or a free choice cue and a subliminal prime congruent or incongruent with respect to the selected response). Here the participant has an actual, but limited kind of control over the action effects, because for each action there is a distinct set of possible color effects, although the specific color is co-determined by the identity of the subliminal prime, which is random with respect to the action and not (completely) accessible to the participants. Participants were asked to assess their sense of control over each color, which again sounds like an unusual and ambiguous task, requiring an interpretation from the side of the participants. This was augmented by the fact that they objectively had equal ability (probability) to produce one of the possible colors through their effort.

The ratings in the original study by Wenke and colleagues (2010) are thus all close to 50 (interpretable as "I don't know" or "I can't judge"). The results show that the ratings are higher when the action is not instructed but up to the participant, which is not that surprising, except for the relatively small effect of only ca. 5 rating-scale points[5]. However, and that seems to be

[5] As far as I can judge from the figures, specific values were not reported

especially clear in the second experiment, the ratings are influenced by the (probably) subliminal action primes, such that when the performed action is congruent with the prime, the rating is slightly higher. The effect is rather small in absolute numbers (mean control rating of ca. 50 for compatible and ca. 48 for incompatible prime-action conditions), but statistically strong[6] (p = .01, d_z = 0.61, achieved power = .78). The interpretation of this finding seems to be much easier than in the other studies, specifically that the action primes led to more effortless action selection and the smoothness of the action was behind the slightly different control ratings. This was probably not reflected consciously, because the authors stated that none of the participants reported having based their ratings on processing fluency, but rather on perceived differences in color frequencies. So it could also be the case that the action priming led to an illusion of frequency differences and this was subsequently used to produce the control ratings, or vice versa, that the priming really led to different sense of action smoothness and this led to the color frequency illusion. Alternatively, people might have been rating perceived affective valence caused by the increased fluency, so that their rating would not mean “I feel like a free agent” but rather “this feels nice” (Grünbaum, 2015). Overall, because there was no apparent objective basis for the participants’ answers to a feeling of control over a color patch, some kind of translation, attribute substitution, is highly likely. It is hard to say what exactly it was participants were rating and we cannot easily conclude that the ratings reflected any experience of agency or of control.

In all the experiments reviewed so far, the amount of actual control afforded to the participants was minimal or none, rating questions were somewhat odd, and the scenarios were rather artificial, far from being ecologically valid. There are experiments that address some of these issues. Notably, Janet Metcalfe and colleagues have conducted a series of studies (Metcalfe, Eich, & Castel, 2010; Metcalfe, Eich, & Miele, 2013; Metcalfe, Van Snellenberg, DeRosse, Balsam, & Malhotra, 2012; Metcalfe & Greene, 2007; Miele, Wager, Mitchell, & Metcalfe, 2011) using a task in a form of a computer game, in which participants have to hit targets while avoiding other items that are all scrolling down a vertical lane, while various aspects of the task are manipulated. For instance, the task can be made easier or harder (scrolling speed, item density), the participant’s control over the mouse control can be lowered by a random “turbulence” of the cursor, the effect (hitting the target which then disappears with a sound) can be less reliable (disappearing only in some cases), and so on. The obtained participants’ ratings can then be analyzed with respect to the various manipulations and to actual performance in the task (objective performance) or to judgements of performance (perceived performance), using regression and mediation analyses, therefore we can

[6] Estimated based on the provided t-statistic and degrees of freedom

more reliably know what factors the control judgments incorporate and how. Therefore, the ratings were not taken at face value as telling us something real about the experience of agency or control, but the nature of the mapping between various factors and, crucially, the possibility of personal interpretation of the rating scale (as either real or perceived performance) were investigated. What is meant by “metacognition of control” was also explained in some detail in the instructions, providing examples like car driving or Ouija boards (Metcalfe & Greene, 2007), although in less detail in the following studies, which only asked for assessing how much participants felt in control, or at least that is stated in the papers.

The description and interpretation of the results from all the studies is too complex to be described in detail. In brief, the first study (Metcalfe & Greene, 2007) found that the control ratings depended to a large degree on the task performance, especially when given salient performance feedback. But more interestingly, that even though diminished objective control over the mouse cursor did not have a big influence on the performance it had an independent influence on the control ratings and inversely, lower speed of the task and “magic” (hitting the target even from larger distance without touching it) had a large influence on performance, but not so much on the control ratings. Therefore, the judgments of agency (JOA)[7] “might have been picking up on the discrepancy between what they, themselves, did or did not do and what the outcome was” (p. 190) and not just on the task performance. The authors considered that the *perceived* performance could have been behind the control ratings and tested this in further experiments by changing the question from being about control to performance (JOP). While the speed and salient feedback manipulations influenced the control and performance judgments in similar ways, the two ratings were differently influenced by the turbulence and magic manipulations. Turbulence led to small decrease in JOP, but big decrease in JOA, and “good” magic led to big increase in JOP but small increase in JOA. In sum, the control ratings were largely but not solely based on how well people were – or thought they were – performing, but also on how well they were able to control the mouse cursor, while the lack of control over the effects of magic was less important to them. In the words of the authors: “people are beautifully sensitive to the kinds of variables that they should be sensitive to in agency monitoring” (p. 195). In another study (Metcalfe et al., 2013), mouse turbulence (“proximal variable”) and reliability of target responding to being hit by the cursor (“distal variable”) were manipulated, and participants reported JOA and JOP on each trial. Analysis of the results showed that “while the proximal variable always had a large direct effect on JOAs, even taking judgments

[7] This term “judgment of agency” was used interchangeably with the term “control” and variations on that term. In their other study (Metcalfe et al., 2013) the term “feeling of agency” is used interchangeably as well.

of performance (JOPs) into account, JOPs completely accounted for the effect of the distal variable" (p. 485). We thus learned from these experiments in some detail on what factors people probably based their control and performance judgments and relations among these factors. It would have been also interesting to ask the participants how they think they made their judgments. From that we could learn valuable information such as whether they were aware of the nature of the relations among the factors and whether all participants understood the task and the questions in the same way or arrived at their judgments in various ways.

It has also been found that children and older adults arrive at their judgments in slightly different ways (Metcalfe et al., 2010). A version of this task was also administered to a group of patients with schizophrenia (Metcalfe et al., 2012). The patients could very well judge their performance, but their agency judgments differed[8] from those of healthy controls in being driven almost exclusively by their JOP and did not incorporate the experimental manipulations, such as the mouse cursor turbulence. The authors claim that the "patient data reveal a lack of metacognition of agency unlike that seen in any group that we have studied to date" (p. 1397) and that "the patients with schizophrenia appeared to be unaware of the presence of turbulence in the mouse controls, or the fact that the response of the cursor was altered by a time lag of up to half a second" (p. 1398). However, as far as I can tell from the report, the patients were not behaviorally tested or simply asked for their awareness of the manipulations. The conclusion about the lack of metacognition of agency does not seem completely warranted, because all that has been shown is that the patients provide approximately the same ratings for both the question about performance and control. We cannot exclude alternative explanations such as that the participants did not understand the instruction to "assess how in control you felt" (p. 1395) properly, or – given that the question is not very usual in our (and perhaps especially their) everyday social interactions and it is thus not completely established what a proper understanding should be – in the same way as healthy controls. Perhaps for them "to be in control" actually means the same as to perform well or perhaps it does not mean much at all to them and basing the reports on performance was as much sense as they could make from the instruction. We cannot be sure – because it was not tested – whether this says something about the patients' metacognition of agency conceptualized and operationalized in other (possibly more informative) ways, e.g., whether they did or did not notice that there was a discrepancy between what they were doing with the mouse and the motion of the cursor (an ability), whether they would say it is unusual or that such things happen to them normally (an

[8] The judgments of agency (in comparison to the judgments of performance and actual performance) were unfortunately not reported, but the composite "agency" contrast scores and regression analysis suggest that the JoAs were very similar to JoPs in the patients.

experience), and so on. It is a strength of the employed experimental design that it can be used to investigate such interesting questions as well.

Studies similar in purpose – testing whether healthy participants and schizophrenia patients can be aware of discrepancies between their movements and sensory feedback – are actually being conducted since the sixties of the previous century. In this class of experiments, based on the so called Nielsen substitution paradigm (Jeannerod, 2006; Nielsen, 1963), subjects often perceived "the effect of the actions of another person, which was substituted for their own" (Jeannerod, 2006, p. 75). In the 1963 study, Torsten Nielsen asked people to draw a straight line while observing their hand through a box equipped with a mirror (unbeknown to the participants) that could on some trials show the experimenter's hand performing the same motion. Participants did not realize it was not their own hand (probably because the deviation was not sufficiently large and it was not made plausible that there could be a second person performing the movements) but adjusted their own hand movements to compensate for observed trajectory deviations[9]. The participants were probably aware that "their" hand moved in a wrong direction but not aware of making the compensatory trajectory adjustments, and some reported a "loss of voluntary control, 'as if driving one's car on an icy road'" (Jeannerod, 2006, p. 76). In several variations of this paradigm (Daprati et al., 1997; Van den Bos & Jeannerod, 2002), people performed a hand movement and observed a movement, which could be either their or the experimenter's and had to discriminate who the actual agent was. Various manipulations are possible in this task, such as temporal and spatial deviations, screen rotations etc. It turns out that people can reliably identify their movement unless the deviations are below a certain threshold, that schizophrenia patients are worse in this task, and that when uncertain, people tend to rather over-attribute the movements to themselves. An important difference from some of the tasks described before is that in this task there is an objective fact of the matter as to who the agent was and there is little ambiguity as to how to understand the task. The task simply explores objective abilities of the participants, namely the ability to recognize their own movements, in a psychophysical fashion, without the need to assume that the recognition is based on any kind of distinctive experience of agency. But note that nothing prevents us from also exploring the experiences of the participants, their reports on how such a task feels, even try to see how the manipulations would map onto other explicit measurements such as the sense of being in control.

[9] Note that the trajectories themselves can be taken as a kind of an implicit measure: of being able to detect a difference between an intended and observed movement at a level which is not necessarily conscious.

The point about the "objectivity" (or, perhaps better, "factuality") can be also summarized like this: The degree to which a task is objective, to which there is a "fact of the matter" to be responded to, can be most easily determined by whether we can *in principle* provide a feedback to each response, saying what the response should have been. In the hand-identification (and similar) studies there is a clear fact of the matter as to whose hand is being displayed and we can therefore truthfully tell the participant: "No, this is not your hand". In the studies by Janet Metcalfe and colleagues, there is *some* fact of the matter as to whether one was in control, *given a specific concept* of control, specific construct behind their question. The authors' concept clearly did not amount to a one correct mapping between all the manipulated factors and the response; therefore there was no correct precise value of the control ratings on any given trial. What mapping would people use was what they were interested in and it is not possible to say whether someone's mapping is correct or not. But their concept presumably did imply that people had objectively less control on trials with mouse-cursor turbulence and it implied perhaps other directional relations as well. The degree to which these relations and the concept of control can be said to be "correct" can depend on experimental instructions but in this case it depended on some assumed ordinary language usage. Unfortunately, the concept of control, or being in control, does not have one common usage and well understood and clearly delineated meaning, and can include such components as mental or physical effort, perceived performance, fluency, number of possible action choices, freedom from instructions, etc. Therefore it is hard to interpret insensitivity to some manipulation (e.g. the turbulence) as a lack of awareness of control, rather than as not having the same concept of control as what the experimenters assume. In the study by Wenke and colleagues (2010), specifically in the second experiment, there was an objective relationship between each combination of action and prime, and the effect. However, is there an objective answer to the question about the sense of control over effects? This question seems to necessarily require some (re-)interpretation, perhaps most meaningfully as being about the probability or frequency of the selected action leading to the displayed effect. This probability was objectively 50 % on all trials and that is also what participants were reporting as their "control" score. In the studies by Sato and Yasuda (2005) there was not any fact of the matter as to who the agent producing a tone *really* was (and it was not possible to provide a feedback about what the correct response should have been), because such a question does not even properly apply in the paradigm used, unless the experimenters' concepts behind the question are radically different from the ordinary language usage and understanding of the explicitly and implicitly involved concepts (e.g., specifying that someone counts as producing a tone only if the same tone happens after the same delay as it used to happen during a previous training, regardless of the actual causal chain of events).

The objectivity criterion is most useful when we want to investigate personal abilities. In that case we can design experiments where correctness-feedback is possible. But when we really want to investigate subjective experiences, we cannot of course (easily) provide such feedback. However, that does not mean that there is no fact of the matter about the nature of subjective experiencing. For instance, there is *some* fact of whether or not (or to what degree) I am feeling pain now, regardless of whether someone else can give me feedback about it. My answer to your question will (most of the time) be valid, because I know what I am being asked about, can detect whether I am in pain and report it. But if I am asked whether I feel control over an effect produced by a computer, I am not sure how to respond. Maybe I feel something that could be described as control, maybe not. Maybe there is no such experience, maybe I do not know what the experience should be like because it does not figure in my ordinary language exchanges, or maybe it is not salient enough for me to detect, individuate, and report it. There is no clear fact to be reported. A report on non-existing, ambiguous, or non-salient experience will not have a high degree of validity as we are not sure what the report is actually about. It will quite possibly involve some rationalization, attribute substitution by the participant.

In conclusion, the less fact of a matter about a required report there is, the more the responses are driven by individual interpretations and the harder it is to interpret the response as validly being about some abilities or processes, such as metacognitive abilities or workings of a forward model, or as about subjective experience. And the more it then reveals the nature of the individual understanding of the instructions and the concepts involved. That can be very interesting, similar to when experimental philosophers try to reveal what concepts such as moral responsibility mean to people (Nahmias, Coates, & Kvaran, 2007; Nahmias, Morris, Nadelhoffer, & Turner, 2005), but also misleading, when this is not the aim of the study. When we use an objective task, we can clearly explain to our participants how the responses should be made and proper understanding of the task is additionally assured by the correctness feedback that we can in principle provide. We can then investigate how the participants' answers deviate from the objectively correct answers under various manipulations, such as in the hand identification paradigm. In other paradigms where there is some objective fact, such as in the study by Wenke and colleagues (2010), we can ask people directly the factual question, i.e., how frequently some effects followed some actions, and observe how the subliminal priming influences such answers. But in that case we should still try to make variations in this objective fact, e.g. changing the frequencies, so that people are not confused that the correct answer is the same throughout the whole experiment and do not attempt individual re-interpretations and rationalizations of the task. Asking about a purported subjective experience of control over a given color in this paradigm does not guarantee (is not only uncertain, but in this case rather unlikely) that the report actually taps into some really felt subjective experience. In paradigms where the objective fact is less well-defined, but we still want to make normative claims e.g. about impaired ability to perceive disturbances in objective control (Metcalfe et al., 2012), we have to make sure that what we mean by control is well understood by the participants. Finally, when we do not want to assume there to be a clear correct basis for answers (e.g., whether an agent was "morally responsible" for an effect) but there is some individual understanding of the concepts, we can investigate what this understanding is, as is the practice in experimental philosophy. But when we really want to gain informative answers about subjective experiences, there are arguably better ways, as discussed further.

2.1.3 Systematic analysis of verbal reports

Behavioral experiments typical of cognitive psychology can tell us a lot about the various abilities pertaining to the sense of agency, such as what factors influence whether I can identify my body parts or the effects of my actions or what counts for me to "be in control" and so on. If we aim to learn about the phenomenal experiences of people in various situations of acting in the world, we can tap into the quantitative but also qualitative methods commonly employed in social psychology and similar disciplines, namely asking people in systematic ways and analyzing such reports. These reports can come from various inventories or structured interviews and can be administered in connection to various tasks and situations. Even though some aspects of our phenomenal agentive experience may be pre-conceptual and not directly available to verbal reports, we can still gain valuable insights into many aspects of personal phenomenal experience. Heterophenomenology (Dennett, 2007) or other types of phenomenological analysis (Gallagher & Zahavi, 2008) can be of use here. It has been suggested that we can tap even into the pre-conceptual experiences through the use of metaphors, so that "an experience (non- or prelinguistic), especially of the prereflective type, becomes progressively conceptualized, i.e. transformed into a conceptual (linguistic) format, in order to be grasped by the reflecting subject, thematized and rendered communicable to others. (...) The metaphor is therefore the first stage of making a prelinguistic or prereflective experience explicitly accessible to oneself and to the other" (Parnas et al., 2005, pp. 237–8). I believe this is especially useful and important in the case of patients with various deficits in agency (the real ability to perform intentional actions) and experience of agency (see section 1.2), to learn more about the nature (and perhaps structure) of their experience, since these are the cases that motivate a large portion of the research on the sense of agency. One good example is the semi-structured

EASE (Examination of anomalous self-experience) interview (Parnas et al., 2005), which can be applied in non-clinical samples as well (Torbet, Schulze, Fiedler, & Reuter, 2015). There is a multitude of inventories for various other experiences, traits, beliefs, and attitudes, which can be informative with respect to experiences of agency, such as schizotypy (Raine, 1991), self-efficacy (Bandura, 1977; Schwarzer & Jerusalem, 1995), locus of control (Levenson, 1981; Rotter, 1966), free will beliefs (Nadelhoffer, Shepard, Nahmias, Sripada, & Ross, 2014; Paulhus & Carey, 2011), mindfulness (K. W. Brown & Ryan, 2003; Walach, Buchheld, Buttenmüller, Kleinknecht, & Schmidt, 2006), and so on.

Similar methods that are behind many of these inventories, i.e., psychometric methods and factor analysis, can be used to study the structure of phenomenal experience in a specific situation. While mere introspection may not be able to capture the essence of the experience, we can ask about the experienced situation in various ways, using various metaphors, and utilize statistical methods to identify some common factors behind the responses. Matthew Longo and colleagues (Longo, Schüür, Kammers, Tsakiris, & Haggard, 2008) exposed 130 volunteers to the rubber hand illusion (see section 1.4.4) in which people often experience the rubber hand as belonging to their body, which can be verified by various objective methods. The authors asked each participant to provide a rating for 27 statements describing various aspects of their experience, e.g., “it seemed like I could have moved the rubber hand if I had wanted” (Longo et al., 2008, p. 984). The 27 statements were selected based on previous free reports of five other participants and some of literature on conceptual distinctions in the sense of agency. Principal components analysis revealed several major components of the experience, depending on whether the rubber hand was stroked synchronously or asynchronously with the real hand, which were interpreted as embodiment of rubber hand, loss of own hand, movement, affect, and deafference and accounted for around 55-60 % of the total variance, depending on the synchronicity condition. Embodiment of rubber hand was further composed of ownership, location, and agency. As the authors suggest, this psychometric approach can “point the way towards an empirically rigorous phenomenology” (p. 979).

Vince Polito and colleagues (Polito, Barnier, & Woody, 2013) have taken the same approach to investigate the experiential structure of agency disruptions in hypnosis. They “identified 24 of the most commonly used words and phrases to describe the subjective experience of agency” in the literature (p. 687), see Table 2-1. Using these terms they composed a 48-item Sense of Agency Rating Scale (SOARS), which they administered to 370 participants after undergoing hypnotic procedure involving ten standard hypnotic suggestions, e.g., that their fingers would feel tightly stuck together. The authors identified two major factors behind the responses, accounting for a little fewer than 50 % of the total variance, which they interpreted as involuntariness ("a change in attributions of personal influence over actions") and effortlessness ("feelings of absorption in the task at hand", p. 689). This two-factor structure was confirmed using a subset of 10 items in a subsequent experiment on 113 more participants who also provided more detailed answers in an open interview. The authors suggest that their "subscales provide an operational conceptualization of sense of agency" and help investigators to "avoid the definitional and conceptual confusion that has traditionally been associated with sense of agency research" (p. 695). I express my sympathies to this project, although the journey ahead of us is still long.

Table 2-1. Words commonly used to describe the subjective experience of agency. *From Longo et al., 2013.*

effortful	planned	under control	surprising
self-generated	easy	compliance	willed
chosen	intentional	deliberate	purposeful
voluntary	caused	inevitable	absorbing
reluctant	responsible	robotic	passive
self initiated	understood	ordinary	unavoidable

2.2 Implicit measures

Obtaining explicit reports from participants can be challenging for a multitude of reasons. As I discuss above, one needs to ask the right questions in the right design, has to pay attention to conceptual and methodological details, cannot take the reports at face value but rather as data in need of careful interpretation, the participants may not be able to access or properly express their phenomenal experience, they may answer a different question than is asked, and may even confabulate their answers. Nisbett and Wilson (1977) in their famous paper "Telling more than we can know" argue that our introspective access to many cognitive *processes* is poor at best and that reports may be based on a priori causal theories of the participants: answering not according to how their processes really operated but according to how they think these processes should operate in the given situation. While they do not express skepticism about access to the *content* of our experience, there are others who do (Schwitzgebel, 2008). In a more applied example, it has been suggested in the field of marketing that if you want to understand consumer preferences, it is better to observe their behavior rather than asking them explicitly (Graves, 2010).

2.2.1 General analysis of the problem

These challenges can motivate us to seek other methods of investigating the sense of agency, the experiences associated with voluntary action, which do not rely on explicit reports. Such methods, which rely on objective data instead, can be called implicit. Authors using implicit measures in the research on the sense of agency for instance argue for their advantage over explicit reports because "[e]xplicit reports of perceived agency are modulated by numerous biases, notably social desirability and cognitive dissonance effects" (Caspar, Christensen, Cleeremans, & Haggard, 2016). However, before we use implicit methods, we should analyze the relations and assumptions behind them in more general terms. Can we infer phenomenal experiences from objective data (not subjective reports) and what conditions need to be fulfilled for that?

Such implicit methods and measures would ideally allow us to infer the nature of an experience E based on an objective measurement M in a task situation S. The S would usually be a task contrast (operant condition versus baseline condition in some task context), M a measurement contrast (an effect, a measurement difference between the two conditions), and E an experiential contrast (the experience which is assumed to be present in the operant condition in addition to the experiences present in the baseline condition[10]). The measure M and experience E both depend on some process (or processes, cognitive, but with a neural basis[11]) P, which is engaged by S. One of probably many possible analyses of this complex situation is depicted in Figure 2-1 and described further.

[10] The E does not need to be some isolated experience or a set of experiences, but rather some general difference in experiencing, because "experiences" of course do not need to be additive in the way we can think our task manipulations are – e.g. active button press can involve experiences which are not present when our finger is passively pressed on the button or does not move at all, but also vice versa – and they do not need to be even discrete or enumerable in any meaningful way.

[11] The process P will be assumed to be a physical (largely neural) causal process which can be described in cognitive or computational terms.

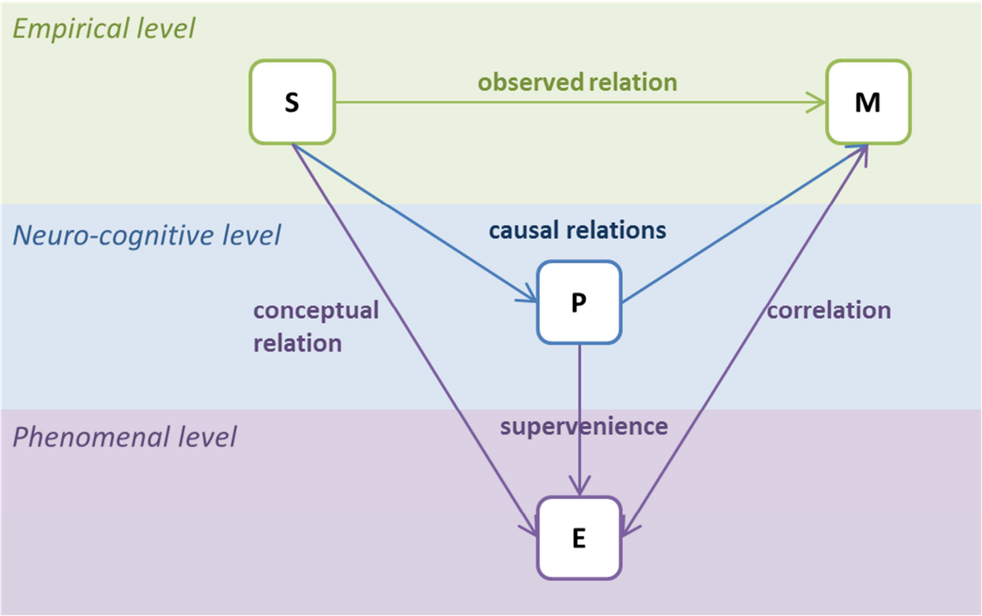

Figure 2-1. A sketch of relations underlying implicit measures of phenomenal experiences. *An experimental situation S leads to observation of measure M, which is produced by a process P and possibly gives rise to an experience E. The direction of the arrows can be read as a functional mapping from set to set, a many-to-one relation, such that e.g. for each S there is some M, but for each M there can be more different S, or equivalently, that if there is a change in M there must have been a change in S, but if there is a change in S, there is not necessarily a change in M.*

The analysis shows that finding and using such implicit measures can be very challenging. We are facing several major issues: First, when we want to establish the implicit measure, we cannot practically arrive at a direct general mapping from any value of M to some E, because having a one-to-one mapping from objective data (e.g. a snapshot of the whole brain) to any subjective experience would amount to having solved the (hard) problem of consciousness. And to get there, we would presumably need to go through the other depicted relations. Therefore, we can start to explore several task situations S and determine the values of M and E in these situations. Recording the values of M for all the S of interest is relatively easy (i.e., finding a mapping S→M). A more difficult matter is determining whether and what E is actually present during S. We can use explicit reports to determine E, but one of the arguments for why we want to establish an implicit measure is that we cannot be sure the report will capture the experience well (Caspar et al., 2016). The report can be seen as just another kind of M, for which reliability, specificity, etc. must be established like for any other kind of M (see below), and in the case of the sense of agency, the experience may be not-well-reportable because it may be pre-reflective, not salient enough, or the participants may simply lack the conceptual repertoire to express it. Nevertheless, I believe that there are many cases in which we can reliably ask for explicit reports and their correlation with the implicit measure can strengthen the relation between the measure and the experience. Alternatively, which is what seems to be the procedure with current implicit measures, we can make an assumption about the E based on the nature of the situation S (or the process P that we assume is engaged by the task situation). We can assume there is something it feels like to be e.g. executing a voluntary as opposed to involuntary goal-directed movement[12], whether the person is reflectively aware of it or not; that is, that there is *some* E during S and just assign this perhaps pre-reflective experience some label. (Similarly, we can think that a process P performs a function that has some feeling to it, e.g., the comparator model for self-attribution of sensory events.) In choosing this label, determining *what* E that is, we are then facing conceptual problems of the same kind as we have discussed above in connection to the search for neural correlates of the sense of agency (see section 1.4), which themselves can be seen as a kind of an implicit measure. Assigning an adequate concept to E depending on the precise nature of the experimental situation requires an a priori conceptual and phenomenological analysis, e.g. separating the sense of agency from the sense of ownership and making other more fine-grained distinctions. Behavioral or neuroimaging measures cannot do this for us. Probably the most popular implicit measure, intentional binding (details in section 2.2.2), has been established in this way, without explicit reports: Difference in S (voluntary vs. involuntary action) conceptually or intuitively thought to correspond to a difference in E (high sense of agency vs. low sense of agency) resulted in a difference in M.

Note that if we were able to precisely assign some E for all S of interest, there would be no need to take the measurements in these S in the future, i.e., no need for the implicit measure at all, because we would know a priori what the experience in a given situation would be. Here (and at other points in this analysis) we have to recognize the problem of granularity, that is, that many of the variables (E, S, P) are not simply continuous, but can be seen as qualitatively discrete, while at the same time coming in degrees. It depends on our conceptualizations what we take to be a different experience (or situation, or process) and what just a different degree (or variation, or state) of the same experience (situation, process). We may therefore a priori assign some *type* of E for some or all S of interest and further perform task manipulations which should (also based on a priori considerations) manipulate the degree of E such that we do not change (much) the type, the concept under which E falls, e.g. providing participants with more or fewer action choices. Similarly, if we have a theory for the process P, the manipulations should still engage the same kind of process P on

[12] That there is some fundamental element of experience common to all the various ways in which we act in everyday lives may be a very strong assumption indeed. It may very well be that there is no special experience of being a free agent at all.

which the E supervenes[13], e.g. a comparator between the expected and actual action effect. We can then record the M under such manipulations. If we then assume that this measure is linearly mapped to the degree of E, we can apply this measure in situations in which it is not a priori obvious what degree E should have to quantify it. The kind of manipulation under which the same concept still applies, the granularity of the concepts, whether the values of the measure map linearly to the degrees (e.g. intensity) of the experience, whether such quantification and comparison of experiences is meaningful in some cases or at all, etcetera, are interesting but not easy questions. For now let us assume that we have established relations S→E and S→M for some S of interest with some granularity and degree of certainty, for instance in probabilistic terms as P(E|S) and P(M|S).

We can also investigate the relations S→P and P→M using classical methods of cognitive psychology and neuroscience and form a theory of how P makes S lead to M. Additionally, we can have a theory about the relation P→E, although this relation is likely to involve a comparatively high degree of uncertainty as the P itself is a matter of supposition. For example we can theorize that there is a process which determines whether I or someone else was the author of some sensory event (S→P). We assume that this process is what gives rise to some experience of what it is like to be an author of some sensory event (P→E). The theory also says that the process leads to attenuation of some measure of perceiving the sensory event (P→M). Whether observing the attenuated measure M actually means that there was a higher experience of causing the event is however far from certain, as is discussed further.

Second, when we want to apply the measure in new (types or only degrees of) situations and infer E from M, i.e., determine what someone experiences when we record some value of M, we are facing the problem of *reverse inference*. The relations that we assume to have established are unidirectional, of the form X→Y, but now we want to reverse some of these relations. Unfortunately, X cannot be deductively inferred from Y, as one Y can be related to many different X. This is a general problem in many areas of science. We can say that the nature of P is under-determined by the data (M and S), because multiple theories about P can in principle account for the data, so we cannot arrive at P via deduction, but rather by abduction, an uncertain "inference to the best explanation". The problem of inferring unobserved experiences from observed data is formally similar to the problem of inferring cognitive processes not only from behavioral but also from neuroimaging data. Russell Poldrack (2006) analyzes the latter problem and shows that the

[13] That some A supervenes on some B means that there cannot be a change in A without a change in B, but not necessarily vice versa.

reverse inference from an activation in area Z in a task comparison A to a cognitive process X – which authors of fMRI papers perform routinely in their discussion sections without much justification – is possible in principle but faces important difficulties. Instead of a deductive inference, we can perform a probabilistic reverse inference using the Bayes' theorem. We need to know several pieces of information to reverse the relation from X→Z to Z→X, to compute the posterior probability P(X|Z) of a process X occurring when we observe activation in area Z: We assume to know the likelihood P(Z|X) of observing the activation when the process is engaged by the task, but we also need to know the prior probability P(X) that the process will be engaged prior to collecting the data (strength of our belief that the process will be engaged by our task comparison), and also the base rate P(Z) of activation in the area regardless of which process is being engaged (selectivity of the measure). Poldrack argues that reverse inference will generally be used to infer the process in a situation in which we do not already assume it being engaged due to the task manipulation (similarly we can often assume the presence of specific type of E during S without the need for any M) and therefore the prior can be quite low. It is important to note that in Bayesian inference we always need to start with a prior belief about what we want to infer and the new measured information can then increase or decrease the strength of our belief. To estimate the selectivity of activation in Broca's area as a measure for the engagement of language function Poldrack uses a database of 3222 fMRI comparisons, including all of the contrasts investigating and not investigating the language function and finding and not finding the activation in Broca's area. In this step he relies not only on a sufficient amount of contrasts in the database but also on ontology (taxonomy) of cognitive processes engaged by various tasks in such a database. Poldrack argues that the lack of a good fine-grained ontology is the most important limitation for performing the reverse inference.

Inference from implicit measures to phenomenal experience is similar to a large degree, but presumably more complicated. We are dealing with a more complicated structure of relations involving not only cognitive processes but also phenomenal experience, where it is hard to estimate the base rates (selectivities) and the priors. We need a good ontology not only of cognitive processes engaged by various tasks but also of phenomenal experiences. Poldrack's analysis dealt with only dichotomous variables and aimed to infer only the presence or absence of a process. In contrast, our variables of interest are often not dichotomous but can take on different types and degrees at the same time (specific value of some measure, intensity of some type of experience), can be multidimensional or structured in a complex way, or perhaps not meaningfully quantifiable at all for some experiences and processes. As has been discussed above, the type of E can be often postulated based on the S alone, because S is usually observed. However, some measures can allow us to determine even S with a good certainty at some granularity due to their selectivity for S. For instance, an EDA (electrodermal activity, also known as galvanic skin response) is a good, although not perfect indicator of some type of emotionally arousing situation, of an engagement of the sympathetic nervous system, and of an arousing experience (Boucsein, 2012). We know a relatively lot about the relations and variables, including the experiencing, in large part due to the salience and therefore reliable reportability of E. If we know the specific nature of S, whether it is e.g. a dangerous or a painful situation, we can often further specify the type of E (based on our folk-phenomenology) as either fear or pain even without an explicit report with some certainty. Once we have determined the concept, or type, for E, at an adequate granularity level, the specific value of M can inform us about the comparative degree of E under various manipulations, e.g. whether some stimulus is more or less fear-eliciting.

Crucially, none of these inferences are completely certain. In general terms, if we observe a change in some measure, we can be sure that there has been some change in the underlying process and the task situation as well. However, the logic of the relations implies that we can never be completely certain that there has also been a change in the experience, even less so what this change was, as the experience supervenes on the process but is not necessarily identical (mapped one-to-one) to it. This is so because while it is hard to deny that an experience depends on some physical processes and when the experience changes the processes must have changed as well, it is not easy to prove that there is some kind of identity between the state of the processes and the experience, such that *any* change in the processes would entail a change in the experience. A simple argument against such a strong ontological position can be that it is generally accepted that most processing in our brains is non-conscious: the processing changes all the time without influencing our experience (even that which influences behavior, e.g. subliminal priming), and conversely, we do not have evidence that there is a special class of (agency-)experience-processes for which *any* change of any of them results in a change of experience. Identity relation is a strong assumption which needs strong justification. In conclusion, we cannot claim any significant certainty for inferences from implicit measures. Investigators using implicit measures of phenomenal experience should be cautious and modest in their interpretations of results and should state and justify their assumptions.

The author also recognizes the complexity of the very large theoretical problem of inferring subjective experiences from objective data and the present treatment should not be taken as complete and without possible problems, but rather as a preliminary analysis of the relations, conditions, and assumptions involved and an illustration of the complexity of the issue.

2.2.2 Sensorimotor implicit measures

Currently, two popular measures are being used as implicitly measuring the sense of agency, which is defined usually as a “feeling that one’s voluntary actions produce external sensory effects” (Yoshie & Haggard, 2013, p. 2028) or in similar ways. These two measures are intentional binding, referring to a contracted reported time between a voluntary action and a sensory effect (Haggard, Clark, & Kalogeras, 2002; Moore & Obhi, 2012) and sensory attenuation, referring to an attenuation of some perceptual and neurophysiological measures for self-produced sensory effects (Blakemore et al., 1998). Both measures are thought to result from low-level sensorimotor processes of action-effect prediction (Waszak, Cardoso-Leite, & Hughes, 2012), possibly from the forward model and prediction-feedback comparator (see section 1.4.1). The general assumption in the literature seems to be that both measures reliably index the sense of agency, taken not only as the ability (process) of identifying the author of a sensory effect, but also as a phenomenal experience. Specifically, these measures are presented as a window into the pre-conceptual, non-reportable feeling of agency (Caspar et al., 2016; Kühn et al., 2011; Moore, Middleton, Haggard, & Fletcher, 2012; Poonian, McFadyen, Ogden, & Cunnington, 2015). We have seen in the previous section that inferring a subjective experience from objective data is far from an easy matter. Moreover, if this experience is supposed to be non-reportable, claims about its presence are hard to falsify.

In comparison, the process (or processes) indexed by the measures may be identified via careful experimental work. It has been proposed that it is not sufficiently clear what this process is, because many studies claiming to investigate intentional binding and sensory attenuation are confounded by manipulating not only action prediction mechanisms, but also “temporal prediction, temporal control, or a general form of identity prediction” (Hughes et al., 2013, p. 133). Hughes and colleagues present a useful conceptual and methodological framework for the separation of these four processes (and there could be other possible processes that their framework does not address). It may for instance be that the implicit measures can be present also in cases in which the participant’s action did not produce a predicted effect and there should thus be no sense of agency for the effect.

With regard to sensory attenuation, Hughes and colleagues report only one study showing perceptual sensory attenuation while controlling for all the confounds identified by them, but only based on a null result in a control experiment with seven participants (Cardoso-Leite, Mamassian, Schütz-Bosbach, & Waszak, 2010). They argue that sensory attenuation can be often driven by temporal prediction (possibly linked to temporal attention) and report studies showing that sensory attenuation can occur completely independently of action if one can predict the identity of the stimulus via other means (Vroomen & Stekelenburg, 2010). And interesting open question is whether all the measures interpreted as showing sensory attenuation – such as signal detection sensitivity, point of subjective equality, BOLD response in various areas, or amplitude of ERP components such as N1, N2, or P3 (Hughes et al., 2013) – reflect the same process.

Regarding intentional binding, which seems to be a particularly popular implicit measure[14], the certainty about its underlying process is not much clearer. On the one hand, many studies are interpreted to support the link between intentional binding and efferent (action-producing) information (Moore & Obhi, 2012). On the other hand, intentional binding has been shown not to be modulated by action prediction mechanisms: mere presence of action (and therefore temporal control) was sufficient to produce the effect (Desantis, Hughes, & Waszak, 2012). Haering and Kiesel (2014) conclude that intentional binding "is not the result of internal prediction due to action-effect bindings, but might rely on higher-order processes" (p. 109). Other studies show that intentional binding can occur even for observed (Poonian & Cunnington, 2013) and involuntary (Dogge, Schaap, Custers, Wegner, & Aarts, 2012) actions and thus needs not be related to agency or intentionality, respectively. Instead, it has been proposed that the "intentional" part in intentional binding is a misnomer, since the effect probably reflects causality in general and therefore should be called temporal or causal binding (Buehner, 2012; Buehner & Humphreys, 2009). Perhaps also the "binding" part is just an artefact of human pattern-seeking caused by the direction of two separate effects, which happen to be directed towards each other. The estimated onset time of the action and of the tone can vary independently (Barlas & Obhi, 2013). Interestingly, in this study by Barlas and Obhi, the overall difference between the two estimates did co-vary with the experimental manipulations, therefore the "binding" interpretation may be justified. Also, the manipulation concerned the number of action options available to participants (one, three, or seven; all associated to the same tone) and resulted in larger binding effect for more options, which could be interpreted as being contrary to the causality inference hypothesis (because the causal link was equally strong, if not even stronger in the one-option condition) and in favor of some degree-of-control hypothesis. More research should elucidate the complex nature of this effect and its underlying process(es).

[14] As of the 14th July, 2016, Google Scholar reports 688 citations of the original study by Haggard, Clark, and Kalogeras, 2002.

All in all, in the words of Dewey and Knoblich (2014), “it is far from clear under what circumstances temporal binding can be considered to be reliable proxy for the [sense of agency]” (p. e110118). The same can be said about sensory attenuation, as our understanding of it is far from clear, although there are several promising theories (Blakemore et al., 1998; H. Brown, Adams, Parees, Edwards, & Friston, 2013; Roussel, Hughes, & Waszak, 2013). Non-selectivity of the measures makes it difficult to infer the cognitive process behind them in a specific new experimental situation, much less the experience, which involves another inferential step (see section 2.2.1). It is very well possible that the process responsible for the temporal binding effect can inform a person about a likely causal relation between two events: the occurrence of an action and a following sensory event (so far almost exclusively a tone). However, it does not necessarily inform the person about who produced the action (self or not-self), because there does not seem to be a special role of the personal efferent information in producing this effect. The detection of causality can of course be one of several cues in determining the agent (section 1.4.3), but the measure by itself does not allow the inference that the person actually self-attributed the sensory event(s). The person could have made a wrong attribution especially in situations similar to experiments by Daniel Wegner and colleagues (section 1.4.2) where it is causally plausible that the person was the agent when in fact someone else was or vice versa. Furthermore, compared to a mere presence of the effect, we are likely to be interested in what a *change* in this measure (between conditions, participants...) signifies. Unfortunately, if we observe a change in the temporal binding measure, it is hard to be sure how the putative causality detection process was changed, whether it e.g. detected stronger or weaker causality, whether determining causality was easier or harder, whether another process needed to be engaged, and so on. Most importantly, even if we grant that there is some process of agent-attribution involved, we do not know whether this process has any phenomenology linked to it and if yes, whether the nature (intensity or kind?) of the phenomenology co-varies with the changes in the process and how (section 2.2.1).

Nevertheless, we can find assumptions in the literature that there is a *selective* link from the effect not only to the process of *agent* identification but also to phenomenology: “subjective contraction of time between an action and its effect only occurs if the patient feels that they are the agent responsible for the action” (Kranick et al., 2013, p. 1110). This study is not an isolated case in relying on such assumptions. A recent study *Coercion Changes the Sense of Agency in the Human Brain* (Caspar et al., 2016) receiving substantial publicity[15] makes explicit claims about changes in experience (referred to as “subjective experience of controlling one’s actions, and, through them,

[15] https://cell.altmetric.com/details/5632859/news

external events.", p. 585, as "low-level subjective feeling of agency", p. 589, or as "sense of being responsible for outcomes of one's actions", p. 590) based on a change in the temporal binding measure: "Using an implicit marker of sense of agency based on time perception, we showed that coercive instructions caused participants to *experience* less agency over the harmful outcomes of their actions" (p. 589, emphasis in original).

In an interesting replication of the famous Milgram's obedience experiments (Milgram, 1963), participants pressed buttons which produced a tone and sometimes gave their fellow co-participants electric shocks, while sometimes being ("freely") able to choose whether to produce the harmful effect and sometimes instructed ("coerced") what to do. In active control condition participants merely pressed buttons and in passive control condition their fingers were pressed by the experimenter, producing a tone in both cases, without any harmful effect. The finding that the temporal estimate between the action and the tone was longer in the coercive condition than in the free condition is certainly an interesting one. Another valuable feature of the study is simultaneous measurement of both the implicit binding measure, ERP measures (specifically auditory N1, often taken as the sensory attenuation implicit measure), and explicit ratings of responsibility, see Table 2-2. Explicit ratings go in the same direction as the temporal binding measure for the free and coerced conditions, although the relationship is not so monotonous when counting in the active and passive control conditions. Interestingly, the N1 amplitudes go in the opposite direction than what is usual in the literature investigating the auditory N1 component as an index of sensory attenuation, where self-produced compared to other-produced tones result in reduced N1 amplitudes (Lange, 2011), while here free and active control conditions resulted in increased N1 amplitudes compared to coerced and passive control conditions. These results could mean that the temporal binding and sensory attenuation as measured by the N1 component may not always be equivalent measures of the same process, although in this case the authors postulate a new common process behind both measures: a "cognitive operation of "distancing", or reducing the linkage between one's own decision-making, action, and outcome" (p. 590). Overall, while the results are certainly very intriguing, one could arguably be more cautious in drawing conclusions that the temporal binding effect signifies "deeply modifie[d]" phenomenal experiencing (and not only certain neuro-cognitive processing, about which a lot is unknown) and that the finding of a change in *experiencing* "may have profound implications for social and legal responsibility" (compared to investigations of an *ability* relevant for ascribing responsibility, such as the "human capacity to control action", which was referred to in the conclusions even though this was actually not the subject of the study; all quotes from p. 590).

Table 2-2. Results of the study by Caspar and colleagues, 2016.

	Interval estimate [ms]	**Interval estimate [ms]**	**Auditory N1 amplitude [µV]**	**Responsibility rating [%]**
Experiment no.	1	2	2	2
Active Control	? (longer than FA, p < .01)	? (similar to FA, p > .9)	-11.1	56
Passive Control	? (similar to CA, p > .6)	? (similar to CA, p > .09)	-10.0	18
Free Action	370	366	-10.7	87
Coerced Action	437	425	-8.2 (or -7.9)	35

2.3 Motivation of own empirical studies

We have seen that the research on the sense of agency needs to take into account important methodological challenges and issues, both when one uses explicit reports and implicit measures. In the present thesis I report the results of two own empirical studies that try to take these issues seriously in their design and inform us, among other things, about the reliability of certain explicit reports and implicit measures.

With respect to explicit measures, we have seen that subjective reports do not need to be always reliable or valid. People may not be able to provide answers to all the possible questions about their agency that we can ask them. Participants may understand the question as being about something else than what the experimenters have in mind, commit an attribute substitution. In such case it is important to carefully explain the concepts that we ask about and how they relate to the task at hand. Or the participants may not be able to answer even a well-understood question at all, because they cannot access such information, they simply do not know, and may commit confabulation. In the first study (chapter 3), with a title *Metacognition of determinants of behavior: Learning to know more that we can tell*, we investigate whether people can report the causes, reasons, or – in a more theoretically neutral manner – determinants of their actions. Can people always tell us why they did what they did? Should we trust their reports? And can they learn to do it better?

The study builds upon previous findings of Simone Kühn and Marcel Brass (2009). They found that under certain conditions people often falsely attribute intentionality or deliberation to actions which were in fact performed automatically, reflexively, that is, they confabulate the reasons for their actions. The study however provoked several new questions, such as: Why were there such large differences between participants? Is it related to some individual traits leading some people to self-attribute events more than other people? For this purpose we employed several questionnaires about personal locus of control, beliefs about free will and about mindfulness. More importantly, could it be that different participants understand the instructions in the same way? People were asked to report on whether their particular action was a result of a decision, but we should not take their understanding of this concept for granted. We therefore explained in detail what specifically it means to make a decision in the task, provided almost unlimited practice and answers to any questions until participants were confident about their understanding and demonstrated it in the practice. Most importantly, Nisbett and Wilson (Nisbett & Wilson, 1977) were famously skeptical about the ability of people to know the causes of own actions, but does this mean that we can never trust such subjective reports or can this ability be trained? Can people learn to know their minds better and give better reports on their agency? To perform such training, we needed to employ a sort of an implicit measure, but not to infer directly the experience of participants but the presence or absence of the deliberation processes. We were able to estimate the probability of an action being deliberate versus automatic based on modelling and classification of reaction times. We also asked for several personal experiences. We asked participants about the feeling of control over their actions, as this is a common question in many experiments, and tested whether people would in fact answer a different question, specifically about perceived difficulty rather than about the actual control over what they can do.

With respect to implicit measures, the biggest question that needs to be explored for any of them is what neural or cognitive processes they actually index. In our second study (chapter 4), with a title *Expect to be distracted: Prediction of salient distractor by action and cue attenuates its interference*, we explored a potential sensorimotor effect related to the implicit measure of sensory attenuation, which could signify the engagement of action-effect prediction mechanisms. These mechanisms are thought to be a low-level cue for self-attribution of sensory events (Synofzik et al., 2008). We thus wanted to infer the presence of action-effect prediction mechanisms, such as the forward model and comparator mechanism and therefore possible self-attribution process, without making any claims about phenomenal experience. To design an experiment disentangling the various processes contributing to the effect, that is, to make sure the effect is due to action-specific and not some other prediction mechanism, we used the methodological framework by Hughes and colleagues (2013). Specifically, in a so-called additional singleton visual search task participants searched for a target while a bright distractor interfered with the task on some trials, causing "attentional capture". In one condition this distractor (its presence and location) was unpredictable, in another it was predictable by an endogenous cue, and in the main condition of interest by a participant's action. Comparison of the attentional capture effect between the action and cue prediction conditions should reveal the contribution of action-specific predictive mechanisms postulated by the comparator model theory (section 1.4.1), while comparison between the cue prediction and baseline conditions should reveal the contribution of a non-action-specific prediction processes, such as postulated by the Bayesian approaches (section 1.4.4). We can thus learn whether our actions *specifically* influence our attention in a way that external events do not and therefore whether such an effect (and the process it indexes) can serve for us as a cue in distinguishing self- and other-produced sensory events.

3 Study 1: Metacognition of Determinants of Behavior

Metacognition of determinants of behavior: Learning to know more that we can tell

Ondřej Havlíček[1,2], Marcel Brass[3], Axel Cleeremans[4], Agnieszka Wykowska[1,5]

[1] Department of Psychology, General and Experimental Psychology Unit, Ludwig-Maximilians-Universität, Munich, Germany

[2] Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität, Munich, Germany

[3] Department of Experimental Psychology, Ghent University, Ghent, Belgium

[4] CRCN, Faculty of Psychology, Université Libre de Bruxelles, Bruxelles, Belgium

[5] Engineering Psychology, Division of Human Work Sciences, Luleå University of Technology, Luleå, Sweden

Author contributions: MB, OH, AW conceived and designed the study. OH performed the research and analyzed data. OH, AW, MB, AC wrote the paper.

3.1 Abstract

There has been a long-standing skepticism towards the possibility and reliability of access to higher-order cognitive processes. People can be for instance substantially wrong about the determinants of their actions: In a previous and present study, participants often cannot distinguish whether their action was intentional, deliberated or automatic, reflexive. The reason for this imperfect metacognition on higher-order processes can dwell in the general unavailability of adequate feedback about the nature of these processes. The unique feature of the present paradigm is that it allows us to determine the processes leading to the actions with reasonable certainty and present participants with exactly such feedback, that is, whether their action was deliberated or reflexive, on a trial by trial basis. Using the fuzzy signal detection theory we found that participants undergoing such feedback-based metacognitive training improved their metacognitive sensitivity significantly more than control participants, while decision bias and other aspects of performance remained unaffected by the training. Individual differences in metacognition that are commonly found thus may not be fixed but rather metacognition may be a trainable skill. We also found a difference between metacognition of first and second order, such that the latter was relatively good even in participants with relatively low first-order metacognition.

Keywords: Metacognition, introspection, agency, decision-making, illusion of choice

3.2 Introduction

In 1977, Richard Nisbett and Timothy Wilson in their article “Telling more than we can know: Verbal reports on mental processes” famously showed that people can be surprisingly poor at carrying out judgments about the basis of their decisions, for instance readily providing reasons for choices that were driven by factors unrelated to their reports. Johansson and colleagues (Johansson et al., 2005) have shown that we can confabulate reasons for choices we did not actually make. More recently, Adam Bear and Paul Bloom (2016) have shown that people tend to self-attribute random outcomes to their own decisions. Originating from the field of social psychology, research on this topic has found new ground in the cognitive sciences, under the more general umbrella term of metacognition.

Metacognition can be characterized as cognition about cognition, or more simply, as the knowledge of one’s own mind. There has been a rapidly growing interest in metacognition across many disciplines — from single-cell animal studies to philosophy of mind — because of the crucial role that metacognition plays in human cognition, such as control and monitoring of mental operations and behavior, self-reflection and self-awareness (Fleming, Dolan, & Frith, 2012). Perhaps surprisingly, there are substantial individual differences in metacognitive skills (e.g., Song et al., 2011), which have been linked to differences in brain anatomy (Fleming, Weil, Nagy, Dolan, & Rees, 2010). However, a crucial question that remains unanswered is whether such differences reflect differences in core brain anatomy or whether they are suggestive of plasticity-related differences and thus metacognition being a trainable skill. Answering this question in the context of metacognition about the determinants of one's actions (compared to the currently more studied perceptual and memory metacognition) was the goal of our study. We wanted to know – in an inversion of the title of Nisbett's and Wilson's article – whether people can learn to know more things about their mental processes that they can then tell us.

Nisbett and Wilson (Nisbett & Wilson, 1977) were skeptical about our access to "higher-order mental processes, such as those involved in evaluation, judgment, problem solving, and the initiation of behavior" (p. 232) and argued that we have rather access to the results of the processes and make inferences about the processes using various information (internal, but mostly external) and a priori causal theories about how the processes should work. Examples of such theories are that the order in which consumer goods are displayed should not influence one's choice (but actually did in their studies) and loud noise should decrease one's rating of a movie (but actually did not). They hypothesized that the problem may lie in the general lack of adequate feedback about the workings of our mind. We therefore cannot improve our theories and use more adequate information in assessing our own cognitive processes.

We thus sought a method which would allow us to do precisely this: uncover to participants of our study the nature of their mental processes. There are several important conditions for such a method if we want to quantify the potential improvement in metacognitive sensitivity to own mental processes: There must be at least two possible causes behind an action, people can be wrong about the causes in a substantial proportion of trials, we can know the real causes with a reasonable certainty on a trial-by-trial basis, and there can be potentially many trials. To our knowledge, there is only one unique paradigm fulfilling such conditions.

Simone Kühn and Marcel Brass (2009) extended the familiar stop-signal paradigm with a decide-signal, such that participants were asked (1, "go") to respond as fast as possible to the occurrence of a stimulus, (2, "stop") to inhibit their response, or (3, "decide") to voluntarily decide whether they want to respond or not. In the third case, participants were asked to make a judgment whether they had voluntarily decided to respond or whether their response was automatic as in condition (1). This design makes it possible to compare objective data (reaction times) and subjective data (metacognitive judgments). Kühn and Brass demonstrated that people often inaccurately thought they had made a decision when in fact their reaction times indicated that their response had simply been reflexive, because a deliberated action involves several more processes than a simple reflexive response, such as detection of a decision signal, inhibition of a reflexive response, deliberation and re-initiation of the response. Their results also revealed substantial individual differences in the proportions of incorrect self-attributions. We capitalized on this pattern to ask whether people can be trained to improve their metacognitive accuracy by providing them with feedback whether their action was likely the result of a deliberation or of an automatic response to the stimulus.

First, we had to make sure that all participants understood what is meant by decision in the present paradigm, namely that to answer “Yes” to the question asking for a presence of decision, one has to be aware of the occurrence of the decide-signal, inhibit the reflexive response, realize this is a decision trial, deliberate whether to reinitiate the response or not based on one’s wishes and the ratio of decisions so far, and possibly reinitiate the response. Answering “No” would mean that the participant responded without even realizing there was a decide-signal. Participants received as much practice and individual explanations as they wanted. Next, all participants performed one session of the main task, largely similar to the design of Kühn and Brass. On a following day, all participants engaged in the same task again, but received a feedback message after each decision trial. In the experimental group, the message was based on their reaction time and performance on the previous day and could read: “You probably did decide.”, “You probably did NOT decide.”, or “Maybe you decided, maybe not.” In the control group, one of the following messages was randomly chosen: “Good.”, “Not that good.”, or “Ok.”. Next, all participants performed a second session of the main task. Last, we administered several questionnaires about the experiment and various constructs. With the help of modelling of the response distributions and fuzzy signal detection theory we were able to estimate the metacognitive sensitivity and bias of each participant and compare how these measures changed as a result of the training. We also analyzed other aspects of performance and the questionnaires.

3.3 Methods

3.3.1 Participants

Overall, 51 participants were tested, but 11 were excluded based on a priori criteria such as performance and understanding of instructions (specified later), to reach a pre-specified number of 40 participants. Twenty subjects were randomly assigned to an experimental group and 20 to a control group.

The age range of the participants was 20 to 33 years (M = 25.4), 38 of them were right-handed and 17 were male. All participants self-reported normal or corrected-to-normal vision. They were paid € 8 per hour or opted to receive a course credit. The experiments were conducted at the Experimental Psychology laboratory of the LMU Munich. All experimental procedures consisted of purely behavioral data collection of healthy adult participants, without involving invasive or potentially dangerous methods. The study was approved by the ethics committee of the LMU Psychology Department, in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Data were stored and analyzed anonymously. All participants provided written, informed consent.

3.3.2 Apparatus and stimuli

Participants were seated in a dimly lit room, wearing sound-attenuating headphones, in front of a LCD monitor (Asus VG248QE, refresh rate 60 Hz, resolution 800×600 pixels), at a viewing distance of approximately 75 cm. A standard keyboard was used to collect responses. Participants were instructed to use their left index finger to press the left arrow key and the right index finger to press the right arrow key.

E-Prime software (Psychology Software Tools Inc., Sharpsburg, PA, version 2.0 Professional) was used to design and present the stimuli. Stimuli consisted of the uppercase letters N, M, V, and W (size 1.3°×1.1° of visual angle) presented in the center of the screen against a grey background (RGB[192, 192, 192], luminance 165 cd/m^2). The letters were presented in black color, but this color could change to either red (RGB[255, 0, 0], luminance 50 cd/m^2) for a “stop” signal or blue (RGB[0, 0, 255], luminance 13 cd/m^2) for a “decide” signal on some trials.

3.3.3 Procedure

3.3.3.1 Choice RT task

The primary task used throughout the whole experiment was a simple 2-alternative-forced-choice letter discrimination task. Each trial started with the presentation of a black fixation cross for 500 ms, followed by a blank screen for another 500 ms, and then by a random one out of the four possible letter stimuli in black color. Participants were asked to respond as fast as possible by pressing the left key for the letters M or V and the right key for the letters N or W. In case the participants pressed the incorrect key or did not respond in the time limit of 2000 ms they received an error feedback screen of a red minus sign on a grey background for 1000 ms. An inter-trial interval (ITI) of 1000 ms followed (blank screen).

3.3.3.2 Stop/Decide-signal task

The main experimental procedure consisted of the same choice RT task as described above with the addition of two possible color signals occurring on 1/3 of the (pseudo-randomly selected) trials. Specifically, in 1/6 of all trials the black letter stimulus would change color to red after a variable delay (stop-signal delay, SSD), signifying a "stop!" signal (De Jong, Coles, Logan, & Gratton, 1990). The participant is supposed to inhibit the ongoing response, i.e., abstain from pressing a button and simply wait for the next trial. In the case of a successful inhibition of the response this trial is labeled as a "stopped" trial and, conversely, as a "failed-to-stop" trial in the case of unsuccessful inhibition, i.e., providing a response. On additional 1/6 of all trials the black letter stimulus could change color to blue after a variable delay (decide-signal delay, DSD), signifying a "decide!" signal. The participant is supposed to inhibit the ongoing response as in the stop trials and to make a decision between resuming the response ("go") or doing nothing and waiting for the next trial ("nogo"). Afterwards a screen appeared asking whether the participant actually decided ("decided" trials) or reacted in a default mode ("failed-to-decide" trials) as in the "failed-to-stop" cases. Participants were instructed to decide about equally often for pressing and not pressing the button, without preparing this decision in advance or having a certain rule in mind. They were only instructed to use their feeling about the approximate ratio of their decisions so far in order to approximately balance their decisions for "going" and "no-going". In the remaining 4/6 of all trials, no color signal was presented, and the task was only to respond with the correct button according to the letter displayed, as fast as possible ("primary response" trials). See Figure 3-1 for a graphical summary of the paradigm and Table 3-1 for an overview of the possible trial types.

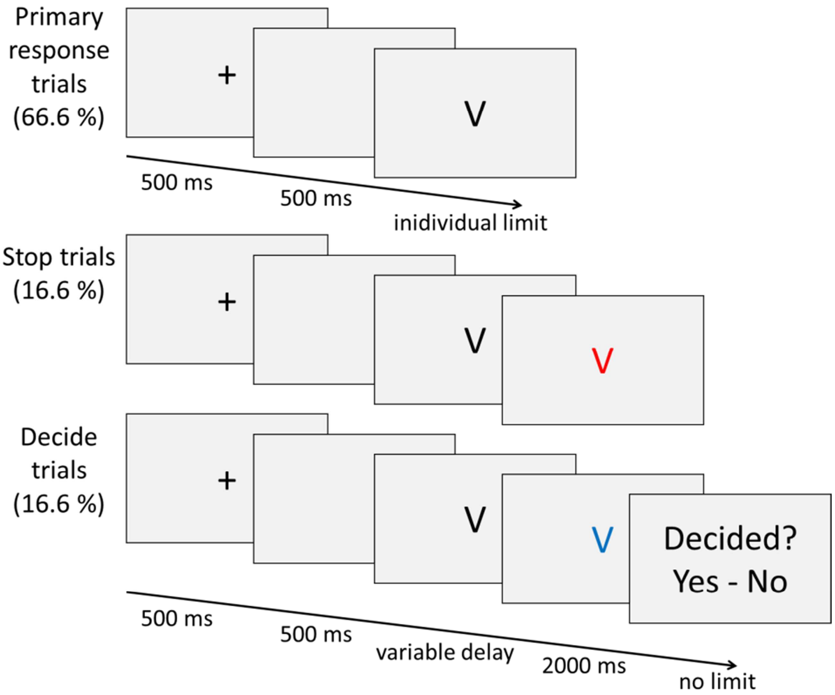

Figure 3-1. Trial sequence in three different conditions. *In the Primary Response condition, participants made speeded discrimination responses to a letter stimulus with a previously established individual time limit. In the Stop condition, a stop signal appeared after a staircase-driven delay, instructing participants not to respond. In the Decide condition, a decide signal instructed participants to make a choice between responding and non-responding and afterwards participants indicated whether what they did was their decision or not.*

Table 3-1. Final type of a trial depending on the condition, response and reported decision.

Condition	Response given?	Decided?	Trial type
PR	Yes	-	Primary response
Stop	Yes	-	Failed-to-stop
	No	-	Stopped
Decision	Yes	Yes	Decided-go
		No	Failed-to-decide-go
	No	Yes	Decided-nogo
		No	Failed-to-decide-nogo

Two separate staircase procedures were used to adjust the stop- and decide-signal delays throughout the experiment in order to maintain approximately 50 % chance of being able to stop and to decide, respectively. The starting value for each of the two staircases was 250 ms and was increased by 20 ms in case of successful stopping or deciding, making the task harder, and conversely, decreased by 20 ms in case of unsuccessful stopping or deciding, making the task easier. The primacy of the choice RT task was emphasized in the instructions. Participants were told that a strategy of waiting for the color signals would be ineffective due to the adaptive staircases, resulting only in ever-longer signal delays.

After each block of trials, participants received a feedback about the ratio of their decisions ("decided-go" vs. "decided-no-go") with a reminder to maintain approximate ratio of 50 : 50. The feedback also included the error rate for color signals (average of their failed-to-stop and failed-to-decide rate), that is, how often they failed to notice the color signal. Together with a reminder that an error rate for colors of 50 % is desirable, they received a message "Try to respond faster." if this error rate was lower than 40 %, "Try to pay more attention to color change." if this error rate was higher than 60 %, and "Your accuracy is quite fine." if the error rate was in between. Additionally, in order to reduce waiting strategies and prolonging the signal delays, another feedback screen was presented, if the mean RT of primary response trials was more than 70 ms longer than mean RT of failed-to-stop trials, consisting of a yellow screen with a red message "Do not wait! Respond faster!" presented for 10 seconds. This annoying feedback has been demonstrated to be very effective against strategic slowing of the primary response (De Jong, Coles, & Logan, 1995) but was implemented only as a safeguard and was not triggered often in our experiment.

3.3.3.3 Metacognitive training task

The main aim of our experiment was to assess how the performance of participants in the Stop/Decide-signal task changes after having received training regarding their decision-making judgments. For this training participants were divided into an experimental and a control group. The metacognitive training task consisted of the same Stop/Decide-signal task as described above with one modification. On each trial in the decide-signal condition, after having responded to the "Decided?" questions, participants received a feedback screen consisting of a short message displayed for 2000 ms. For the experimental group, these messages could be (1) "You probably did decide.", (2) "You probably did NOT decide.", and (3) "Maybe you decided, maybe not.". In the case of trials on which participants abstained from responding the first message was always displayed. For the trials on which the participants actually did respond, the message was selected according to the reaction time that it took the participant to respond. The first message would be displayed if the RT was longer than a "late threshold", the second message would be displayed if the RT was shorter than an "early threshold", and the third message would be displayed for RTs in between the thresholds. The two thresholds were determined for each subject based on their primary response RTs in the first session of the Stop/Decide-signal task, as described further. For the control group, the messages could be (1) “Good.”, (2) “Not that good.”, and (3) “Ok.” and were selected randomly, regardless of the participants’ behavior. These messages were selected to control for the effect of emotional valence of the feedback (and possible associated effects of reward, alerting, or causing unspecific reflections on the task at hand), but without being informative with regard to the decision-making. The participants in either group were not instructed about the meaning of these feedbacks.

3.3.3.4 Overall structure of the experiment

The experiment took place on two consecutive days. On the first day, participants received detailed instructions and performed two blocks of a simple choice RT task, with 50 trials in each block. The first block served only the purpose of training the stimulus-response mapping and the second one was additionally used to determine the individual response time limit for the subsequent Stop/Decide-signal tasks. This limit was computed as the mean reaction time on correct trials plus two standard deviations and was employed in order to minimize waiting strategies in the Stop/Decide-signal task, i.e., postponing the response to see whether a color signal would appear on a given trial (Chen, Muggleton, Tzeng, Hung, & Juan, 2009). The average response time limit was 809 ms. Next, participants practiced four blocks of 30 trials of the whole Stop/Decide-signal task. The instructions for the task and the meaning of the feedback screens were thoroughly (re-)explained to the participants during this time. The average stop- and decide-signal delays from this practice session were saved and used as the initial values for the next part of the experiment, that is, a first session of the experiment proper. This practice of transferring the average signal delays as the initial values for the next experimental task was used throughout the whole experiment. At the end of the first day, participants performed the first session of the main experimental task, that is, the Stop/Decide-signal task, consisting of 12 blocks of 60 trials (i.e., 720 trials in total).

On the second day, participants performed six blocks of 60 trials (i.e., 360 trials) of the metacognitive training task. For the experimental group, the two RT thresholds used to select the feedback message for the decision trials were based on the RTs of the individual participants in the first session of the Stop/Decide-signal task. The reason for this was that the primary response distribution provides us with the best estimate of the time it takes the participant to simply respond to the primary choice RT task, without involving any decision-making processes. The early and late thresholds were chosen to be 0.6049 and 0.9627 quantiles, respectively, of the individual PR RT distribution. These quantiles were determined based on pilot data as on average resulting in a low chance of misclassification of the reaction times as belonging to the wrong of the two ("early" or "late") decision RT classes, see Supplementary information. After the metacognitive training, the second session of the Stop/Decide-signal task followed, with the same properties as the first one.

In the end, participants filled-in several questionnaires. The first one was a custom questionnaire asking the participants several questions regarding their perception of and feelings about their performance in the tasks. On a scale from 1 to 7 (with the two anchors labeled by specific descriptions) we asked about the general perceived difficulty of the whole experiment, how often they were aware of waiting for the color signals, how well they understood what they were supposed to do in the decision condition, how hard it was to make proper decisions, how sure they were about their responses to the Decided? question, and for most of these items (all except for the waiting question) an additional sub-item asking for a comparative judgment with regard to the previous session (e.g. whether they think they were sure about their responses in the last session to a higher, lower, or the same degree as in the first session, on a scale from 1 to 7). We also asked how much they felt in control over their actions in the three different conditions (i.e., three separate items), and how difficult the three conditions were for them (three items as well). Additional open questions asked about their strategies and what they thought about the feedback messages in the metacognitive training session. We started administering several further questionnaires only at later stage of the data collection. These consisted of the Mindful Attention Awareness Scale (MAAS) (K. W. Brown & Ryan, 2003), obtained from 30 participants; Free Will and Determinism Plus (FAD+) scales (Paulhus & Carey, 2011), 30 participants; and locus of control (LOC) was measured using the Internal, Powerful Others, and Chance (IPC) scales (Levenson, 1981), 23 participants. Debriefing about the purpose of the experiment and the nature of the metacognitive training was provided to the participants. In total, the experiment took about 3-4 hours per participant.

3.3.4 Analysis

3.3.4.1 Participant exclusion criteria

In order to assure that participants understood the task and were able to perform it according to the instructions, several criteria were chosen a priori and had to be met in both sessions of the main task. The criterial measures were computed and checked before computing other results. Even though we employed several methods for preventing waiting strategies, we were able to measure how much participants actually relied on them in several ways: The ratio of PR trials on which participants failed to respond within a time limit was required to be lower than 25 % (met by all participants; reported further are statistics for the included participants over both sessions: M = 5.52 %, SD = 4.78 %). Furthermore, the failed-to-stop and failed-to-decide rates were required to be in a range between 25 and 75 % (not met by four participants; failed-to-stop rate: M = 50.65 %, SD = 4.64 %, failed-to-decide rate: M = 44.97 %, SD = 4.08 %). The staircase procedures should maintain these rates close to 50 %, therefore low values of these rates indicate waiting strategies and high values indicate impulsiveness or a lack of attention. Impulsiveness or a lack of attention can also be measured by the ratio of incorrect responses to the letter discrimination task in the PR condition, which was required to be below 25 % (not met by two participants; M = 7.59 % , SD = 5.43 %). In the Decide condition participants were instructed to decide for pressing the button (decided-go) in approximately 50 % of the cases and for waiting (decided-nogo) in the remaining 50 %. For inclusion in the analysis, the ratio of decided-go trials was required to be in a range between 25 and 75 % (not met by five participants; M = 52.32 %, SD = 9.65 %). Finally, we measured how often participants said they had not decided in cases when they did not press any button in the Decide condition (failed-to-decide-nogo ratio). This happened very rarely and high values of this measure (upper limit was set to 25 %) would indicate lack of understanding of the task, which was subsequently confirmed by asking two excluded participants with failed-to-decide-nogo ratios of 51 and 61 % (for the included participants M = 1.33 %, SD = 3.21 %). Some participants did not meet more than one criterion. In total, eleven participants were excluded (replaced by new ones) because of not having met the criteria.

3.3.4.2 Trial exclusion criteria

Several types of records were removed from the analyses. We removed trials with incorrect response to the letter discrimination task in any of the conditions (6.85 % of all trials). We also removed outliers in the primary response condition (defined as RT outside of the range of mean plus or minus 2.5 standard deviations of logarithmically transformed RTs, per subject and session), to prevent a long tail of the PR data bias a classifier constructed based on the PR data in later stages of the analysis (1.05 % of correct PR trials).

3.3.4.3 Modelling the decided-go RT distribution

As expected based on the previous study by Kühn and Brass (2009) and our pilot data (see Supplementary information), visual inspection of the reaction times of the decided-go trials (of all the participants taken together) revealed a bimodal distribution (Figure 3-2). The authors concluded that the bimodal shape of the distribution is an evidence of two distinct underlying processes. To establish whether we can make such a conclusion based on the present data, it is important to statistically evaluate the nature of the decided-go RT distribution by the following procedure: The decided-go RT data were randomly divided into two halves. We used the maximum likelihood estimation (MLE) method to fit several distributions to the first half of the data: normal, log-normal, bimodal normal, and bimodal log-normal. We then performed a one-sample Kolmogorov-Smirnov (KS) test to test the null hypothesis that the second half of the data was drawn from the fitted distributions. This whole process was repeated 1000 times. The average test statistics and corresponding p-values were used to evaluate the goodness of fit. Out of the distributions for which the KS test did not reject the null hypothesis, the best model was chosen based on the Akaike information criterion (AIC) of a model fitted to the complete (undivided) data. Based on this procedure, we selected the bimodal log-normal distribution as the best model of the decided-go trials (see Supplementary information).

This bimodal distribution strongly suggests the existence of two different processes underlying responses in the decision trials. One process is responsible for the short reaction times ("early decided-go"), in the range of PR and failed-to-stop responses, thus very likely being the same process of simple responses to the letter discrimination task, without additional recognition of the blue decision signal, response inhibition, deliberation whether to respond on the given trial, and re-initiation of the response. Second process is responsible for the long reaction times ("late decided-go"), probably initiated some time after the initiation of the first process, namely after the presentation of the blue decision signal, and reflecting the processing of this decision signal and deliberation about whether to respond on the trial. In the perspective of the signal decision theory (SDT, Green & Swets, 1966), participants were asked to detect and report the presence of the second type of a process (a mental "signal") when presented with the "Decided?" question. Answering "Yes" when the response was driven by a process of the second type can thus be conceptualized as a hit, and conversely, answering "Yes" when the response was driven only by a process of the first type can be conceptualized as a false alarm.

Crucially, the bimodal distribution of the response RTs allows us to determine with a certain degree of certainty whether the process of the second type was behind the response or not, and therefore classify the decided-go trials as truly deliberated (late decided-gos, i.e., hits) or as falsely attributed decisions, when in fact the process of deliberation did not have time to occur (early-decided-gos, i.e., false alarms).

In order to assess the effect of our metacognitive training manipulation, we need to make a choice over (i) dependent measure(s), reflecting how well people can distinguish these two processes and (ii) the specific way of classifying the decision trials into early- and late-decided-go classes.

3.3.4.4 Dependent measures

In our study we are mainly interested in quantifying the ability of participants to discriminate between reflexive and deliberated-upon actions and how this ability changes in the experimental group as a result of the metacognitive training and in the control group. Given that we can classify trials into early- and late-decided-go classes, a relatively simple measure of this ability is a ratio of the number of trials in these two classes. This “early/late ratio”, or, “false-alarm/hit ratio”, is also what Kühn and Brass decided for and can be calculated as the number of early decided-go trials divided by the collective number of early- and late-decided-go trials.

However, with respect to our research question, this measure cannot distinguish whether a potential effect of the metacognitive training is due a change in the sensitivity to the mental processes (ability to tell apart real decisions and reflexive responses) or due to a simple shift in the decision criterion (e.g., setting a different internal RT threshold for calling something a decision). In other words, we can obtain more information than the early-late ratio can tell us by computing the signal decision theory measures of sensitivity and bias. To compute these measures one additionally needs the number of misses and correct rejections. These can be best estimated in the same way as the hits and false alarms, by classifying the failed-to-decide-go trials (those trials in which there was a blue decision signal, participants responded by pressing a button, but indicated they did not make a decision) into an early and late class. From the distribution of the decided-go trials we know the distributions of (the finishing time of) the first and the second process and can use this information to classify the failed-to-decide-go trials as either misses or correct rejections.[16]

In sum, we can observe our effects of interest via three main dependent measures: the early/late ratio, sensitivity (d-prime) and bias (criterion), out of which we prefer the latter two.

[16] While it is conceptually unproblematic to classify early failed-to-decide-go trials as correct rejections (the short range of RTs is not enough time for the deliberation process), it is more complicated with the late failed-to-decided-go trials. One could argue that the prolonged response time (beyond the time needed for the letter discrimination) is not due to a presence of deliberation over whether or not to respond, but due to other processes like absent-mindedness. At any rate, the number of such late-response failed-to-decided-go trials was generally very small, as participants seem to be well able not to miss the presence of the deliberation process. This fact of the occurrence of zero numbers of misses in some participants requires the use of the log-linear correction of each of the four cell frequencies in the SDT contingency table, which should produce a minimal bias for the current numbers of trials and actually obtained levels of sensitivity (Hautus, 1995) and which some authors recommend to apply routinely, irrespective of the occurrence of extreme values (Snodgrass & Corwin, 1988).

3.3.4.5 Collective-level classification

It is important to make clear at this point that the model fitting and related operations have been so far performed on the collective level, that is, on all the RT data from all participants, as there are not enough trials per participant to fit the model to individual data. We will therefore first address how the decision trials can be classified at the collective level, as it was done in the original study by Kühn and Brass, and extend the approach to individual data later.

There are in principle two general ways of classifying the decision trials (decided-go and failed-to-decided-go) into two classes. The first way, used by Kühn and Brass, is a crisp classification based on two threshold RT values, such that trials with RT below the early-threshold are taken as belonging to the early-decided-go class, trials with RT above the late-threshold as belonging to the late-decided-go class and trials in between the two thresholds are left out of the analysis. The two threshold RT values can be selected based on e.g. the criterion of minimizing incorrect classification, in our case such that there is only a 1% chance that the trial belongs to the other RT distribution. That is, the early-threshold is computed as the 0.01 quantile of the modelled late-decided-go distribution, and the late-threshold as the 0.99 quantile of the modelled early-decided-go distribution. See Figure 3-2.

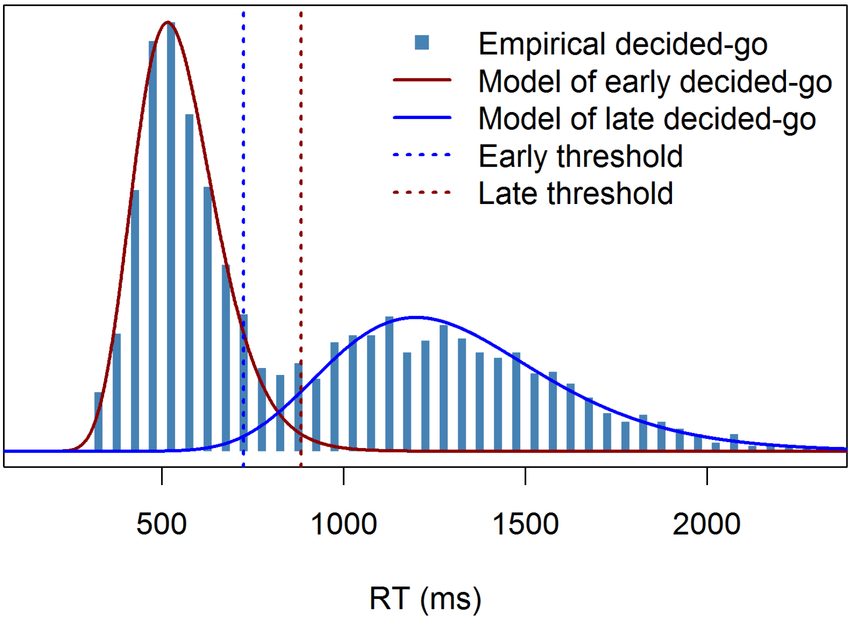

Figure 3-2. Crisp classification of decided-go data. *Depicted are a histogram of empirical reaction times of the decided-go trials, fitted models of an early and late distribution of the trials, and two crisp cut-off thresholds for assigning trials into one of the two distributions. The y-axis is in arbitrary units.*

Crisp classification suffers from two shortcomings because it disregards the actual RT value of a given trial. First shortcoming is statistical, that is, this method does not take into account the fact that the actual RT value carries some information about the probability of belonging to each of the two distributions, but rather classifies trials near a threshold in the same way as a trial on the same side but much farther away from the threshold. Second shortcoming is conceptual, specifically, that the method does not take into account that what we call as deliberation or decision-making can come in degrees. While some decisions can take a long time, explicitly reflecting on reasons for the possible courses of action, there are also “snap decisions”, performed in a short time without much explicit deliberation, which nevertheless conceptually qualify as decisions, although to a lesser degree. For these reasons we prefer a fuzzy classification approach, which assigns each trial with class-membership degrees, depending on the specific RT value.

Instead of two cut-off thresholds, this approach works with two classification functions (also known as membership functions or classifiers; these terms will be used interchangeably), one for each class. Note that this in general allows one trial to belong to a certain degree to both classes. We can compute the classification functions simply as the relative probability that a trial belongs to a given one of the two estimated RT distributions by dividing the estimated probability density function (PDF) by the sum of both of these density functions. Specifically, the early classification function is equal to the early-decided-go PDF divided by the sum of the early- and late-decided-go PDFs. This classifier has a shape of a sigmoid function and takes on values between zero and one, assigning high membership degrees to short RTs and low membership degrees to long RTs, see

Figure **3-3**. The late classification function is constructed in the same way and is a complement to the early classification functions, such that for any RT value the membership degrees of a trial belonging to the early and to the late class sum up to one. Additional benefit of this approach is that the sum of all these membership degrees is equal to the number of trials. We can therefore easily obtain the SDT contingency table cell frequencies by simply summing these membership degrees of the early and late classes for the decided-go trials (yielding the number of false alarms and hits) and for the failed-to-decided-go trials (yielding the number of correct rejections and misses). This approach is in essence a case of the fuzzy signal detection theory (Parasuraman, Masalonis, & Hancock, 2000).

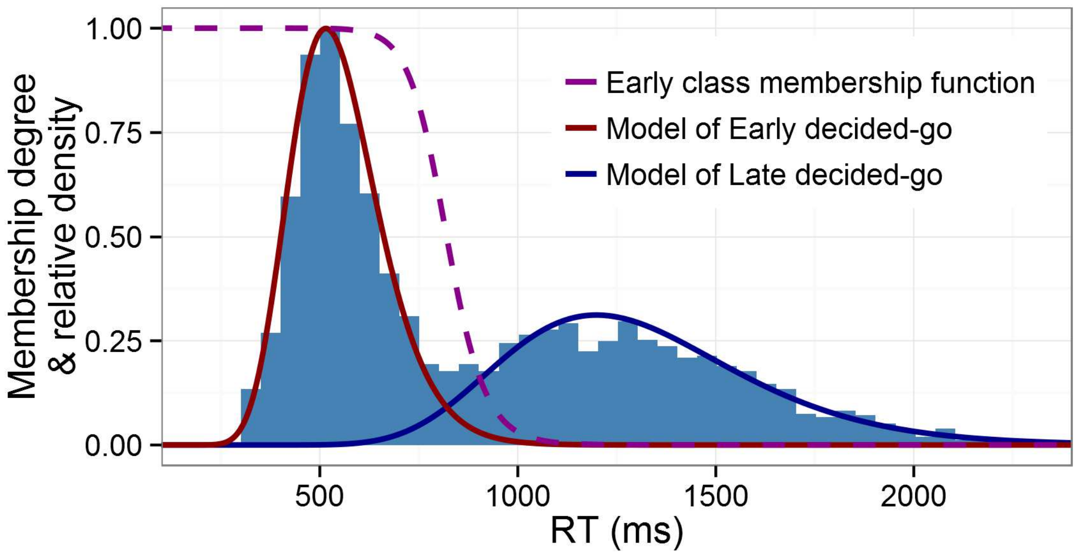

Figure 3-3. Fuzzy classification of decided-go data. *Depicted are a histogram of empirical reaction times of the decided-go trials, fitted models of an early and late distribution of the trials, and a fuzzy membership function for the early distribution.*

In sum, we can use a crisp or a fuzzy classification approach, out of which the fuzzy approach is the one we prefer.

3.3.4.6 Individual-level classification

In this study we aim to obtain the dependent variables not only at the collective level but also for each individual participant. This aim faces obvious challenges. It is possible to estimate the early and late RT distributions only at the collective level, because there are not enough decided-go trials at the individual level (on average 30.6 trials per participant and session) and because not all participants have both early and late trials. We cannot directly apply the thresholds or classification functions derived at the collective level to the individual data to obtain the individual measures, because each participant has a different general response speed and response variance. We propose that the individual pattern of responding can be best taken into account using the information provided by the individual PR RT distribution (see Supplementary information). This distribution can be modelled easily thanks to the rather large amount of PR trials (on average 419 trials per participant and session).

Individual crisp classification can be achieved by expressing the two collective-level RT threshold values as quantile values of the collective-level PR distribution (modelled as a lognormal distribution). For example, for the data from all subjects and both experimental sessions taken together, the early-threshold has a RT value of 725 ms, which corresponds to the 0.9598th quantile (or z-score of 1.749) of the collective-level PR distribution. For the late threshold it is 883 ms, 0.9964th quantile (z-score of 2.688). These collective-level quantiles can then be applied at the individual level, to the individual PR distributions, to obtain the individual threshold RT values. That means that in this case an average person needs a response time at least 2.688 standard deviations longer than the mean of his or her (log-transformed) PR distribution to classify this trial as a hit and shorter than 1.749 SDs from the mean to classify it as a false alarm.

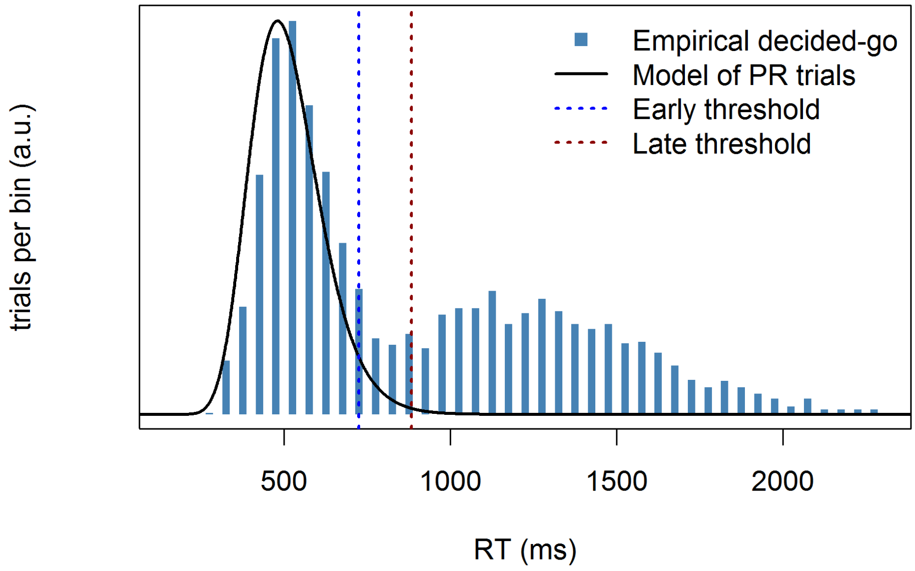

Figure 3-4. Crisp classification based on PR trials. *Collective-level early and late thresholds can be expressed as quantiles of the collective-level PR distribution to construct individual-level thresholds.*

In order to perform individual fuzzy classification, we need to take a similar approach, constructing for each participant a sigmoid-shaped classification functions based on the collective-level classification functions, but adjusted for individual response speed and variance. We can construct such a classifier as a cumulative distribution function (CDF) for a lognormal distribution, with two parameters: a mean (signifying an RT value at which it is equally likely for a trial to belong to either of the two classes, reflecting individual response speed) and a standard deviation (SD, signifying a slope of the sigmoid function, the degree with which the certainty of classification increases when moving away from the mean, reflecting individual response variance). The individual mean can be determined using the same way as the crisp classification thresholds: Finding the RT value at which a trial is equally likely to belong to either of the two classes at the collective level, expressing this RT value as a quantile of the collective-level PR distribution (q = 0.9903, z = 2.3374, for the data from all subjects and both sessions), and determining what RT value corresponds to this quantile value of the individual PR distribution. For the individual SD value one could think about using simply the SD of the individual PR distribution, but the resulting individual classification function would not on average match the slope and thus the overall shape of the collective-level, on-average-optimal classification function. Therefore the SD needs to be adjusted by a constant factor to achieve this match. We can find this constant by a simple least-squares optimization: We construct a lognormal CDF at the collective-level, with the mean equal to the RT value with equal likelihood of belonging to both classes, and the SD equal to the SD of the collective-level PR distribution times the unknown constant factor. We then search for such a value of this factor, which minimizes the sum of squared differences between this CDF and the collective-level classifier, over the range of the data we want to classify. The resulting fit is very close (Figure 3-5), suggesting that a lognormal CDF is a good proxy for the optimal classification function at the collective level, and because its parameters can be easily computed for each participant, can be applied for classification at the individual level. The individual classification functions may be biased (e.g. due to unusual shapes of the individual PR distribution) for some participants, because the function parameters are based on collective-level data, but this fact should also mean that the biases should be balanced out on average.

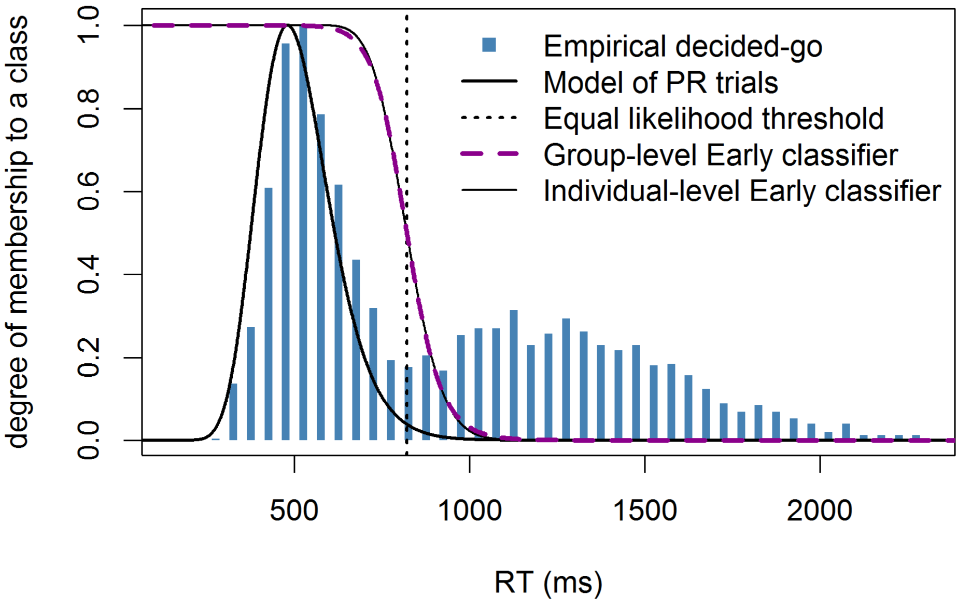

__Figure 3-5. Fuzzy classification based on PR trials.__ Parameters (equal likelihood threshold and slope) of a collective-level fuzzy classifier of the early distribution can be expressed using the collective-level PR distribution to construct individual-level classifiers.

3.3.4.7 Overall analysis procedure

The analysis can be summarized briefly like this: Participants were excluded on the criteria of performing the tasks according to instructions. Outlier trials were excluded only in the primary response condition, to prevent long tails from biasing the individual classifications which are based on the PR distribution. Data from all participants and both sessions were used for modelling the data and determining parameters for subsequent individual classification. We decided to use data from all participants and both sessions as opposed to e.g. modelling the two sessions or the two groups of participants separately as it is the simplest option and provides us with more data for model estimation. Decided-go and failed-to-decide-go trials were classified into early and late classes in order to compute the main outcome measures (sensitivity and bias). There are two ways of performing the classification: a crisp classification based on two thresholds and a fuzzy classification based on membership functions. We performed a fuzzy classification of the trials so that for each trial there is a degree of membership to the early and the late class. The membership function was computed on the collective-level data and has two parameters: an equal membership probability RT value and a slope. These two collective-level parameters were transferred to the individual level by expressing them in terms of the PR distribution. Specifically, the equal probability values were based on PR quantiles and the classifier's slope was based on the PR standard deviation. The classification yields four classes of trials (hits, false alarms, correct rejections, and misses). Individual trial frequencies in these classes were used to compute our dependent measures: sensitivity and bias. Log-linear corrections of the hit-rate and false-alarm rate were applied.

The influence of metacognitive training on the dependent measures is to be evaluated using mixed-design analysis of variance (ANOVA, within-subjects factor "session number" and between-subjects factor "participant's group"), which are to be followed by paired-samples t-tests investigating effects within the two groups (determining whether there was a change in the dependent measures in the individual two groups after the training session), and Welch's t-test investigating effects between the groups (determining whether the change in the experimental group was different than the change in the control group). Questionnaire data (Likert items) were handled as ordinal data and their association with other measures was investigated using Spearman's rank-order correlation.

3.4 Results

3.4.1 Overview of response times

Distribution of reaction times per the four trial types together for all participants and sessions can be seen in Figure 3-6; means of median reaction times per trial type, session, and group are presented in Table 3-2. Evaluating differences in median reaction times between the various trial types, individually for all four combinations of session and group, we found the PR trials to be significantly longer than the failed-to-stop trials (all four *t*[19] > 3.35, p < .004) and the failed-to-decide-go trials (all *t*[19] > 2.83, p < .011), which could be attributed to waiting for the color signals in the PR condition, although this waiting was on average very short, around 20-40 ms, showing that our design was effective at preventing waiting strategies. There was no significant difference between the failed-to-stop and the failed-to-decide-go median RTs (all *t*[19] < 0.99, p > .337), which is in line with the hypothesis that both types of trials were generated by the same process, namely reflexive response to the letter stimuli.

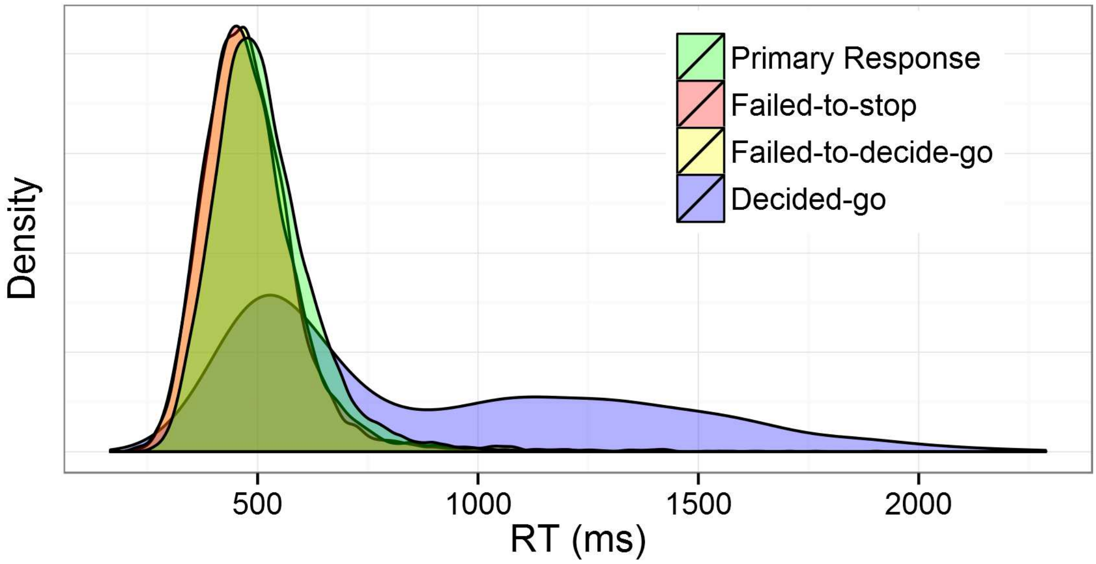

Figure 3-6. Reaction time distributions (kernel density estimates) per trial type for all participants and sessions.

Table 3-2. Means (and standard deviations) of individual median reaction times.

	Experimental group		Control group	
Trial type	Session 1	Session 2	Session 1	Session 2
Primary Response	513 (57)	503 (68)	510 (58)	494 (53)

Failed-to-stop	477 (44)	466 (56)	470 (53)	467 (52)
Failed-to-decide-go	479 (56)	472 (65)	477 (54)	472 (60)
Decided-go	738 (243)	983 (385)	851 (374)	909 (376)

Inspecting the RT distributions of decided-go trials (Figure 3-7) one can observe a large number of trials in the range of primary responses in all the four combinations of group and session. There is a clear decrease in the amount of these early responses and increase of the late responses in the experimental group between the sessions. However there is a similar (although smaller) trend visible for the control group as well. This aggregate perspective thus does not allow conclusions about the effect of the metacognitive training. To be able to make such conclusions we need to perform individual-level analysis.

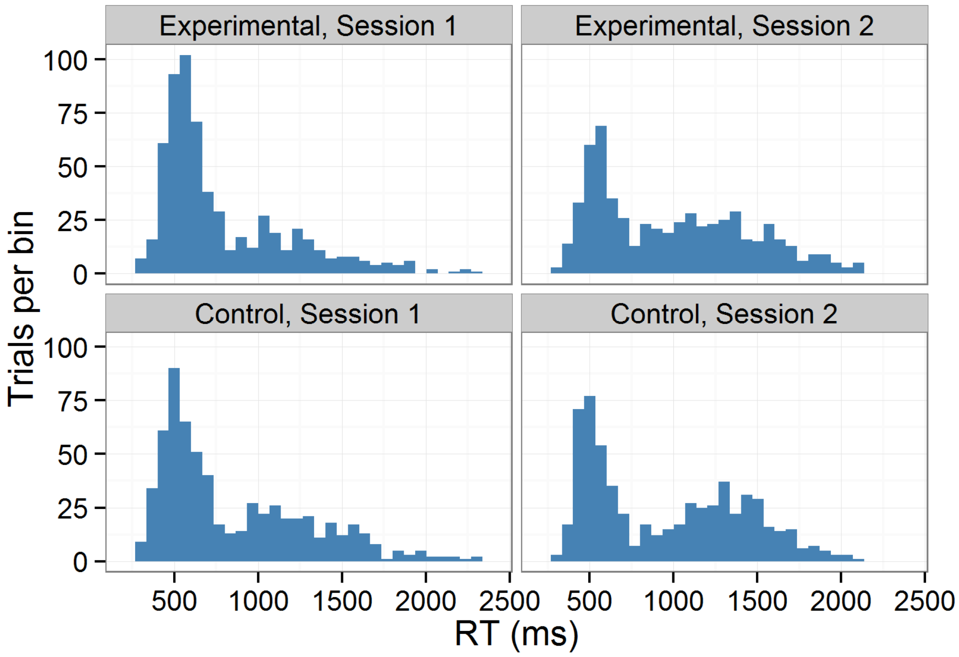

Figure 3-7. Reaction time histograms of decided-go trials per session and group.

For additional reaction times and other analyses see Supplementary information (section 3.7).

3.4.2 Metacognitive training effects

The fuzzy SDT analysis described above allowed us to compute a d-prime (sensitivity) and a criterion (bias) value per each participant in each session of the experiment. ANOVA conducted on the d-prime measure, with a within-subjects factor "session" and a between-subjects factor "group", revealed a significant main effect of session, with d-prime being higher in session two (M = 2.32, *SD* = 1.25) than in session one (*M* = 1.84, *SD* = 0.98; *F*[1, 38] = 9.31, p = .004, η_G^2 = 0.0451, η_p^2 = .197), no significant effect of group (*F*[1, 38] = 0.158, p = .693, η_G^2 = 0.0034, η_p^2 = .004), but a significant interaction of session and group, *F*(1, 38) = 4.63, p = .038, η_G^2 = 0.0229, η_p^2 = .108. Follow-up analysis of the main effect of session with dependent *t*-tests showed that the mean change in d-prime between sessions (95% CI = [0.30;1.31]) was highly significant for the experimental group (*t*[19] = 3.35, *p* = .003; Hedges' *g* = 0.77, 95% CI [0.26;1.32]), but the change (95% CI = [-0.27;0.55]) was not significant for the control group (*t*[19] = 0.71, *p* = .484; Hedges' *g* = 0.11, 95% CI [-0.2;0.43]). Follow-up analysis on the interaction showed that the mean d-prime change in the experimental group (*M* = 0.81, *SD* = 1.08, *n* = 20) was greater than in the control group (*M* = 0.14, *SD* = 0.88, *n* = 20). The mean difference in d-prime change between the two groups (*M* = 0.67, 95% CI = [0.04;1.3]) was significant as determined by Welch's *t*-test, *t*(36.49) = 2.15, *p* = .038; Hedges' *g* = 0.67, 95% CI [0.04;1.31]. The observed data are surprising under the assumption of a null hypothesis that the training did not have an effect on the metacognitive sensitivity. See Figure 3-8 for aggregate and Figure 3-9 for individual perspective.

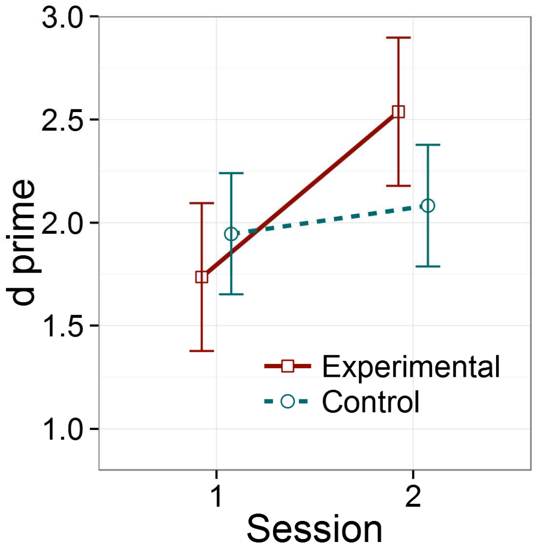

Figure 3-8. Mean metacognitive sensitivity per session and group. *Error bars depict 95% CI for the means, corrected for dependence in measurements (Morey, 2008).*

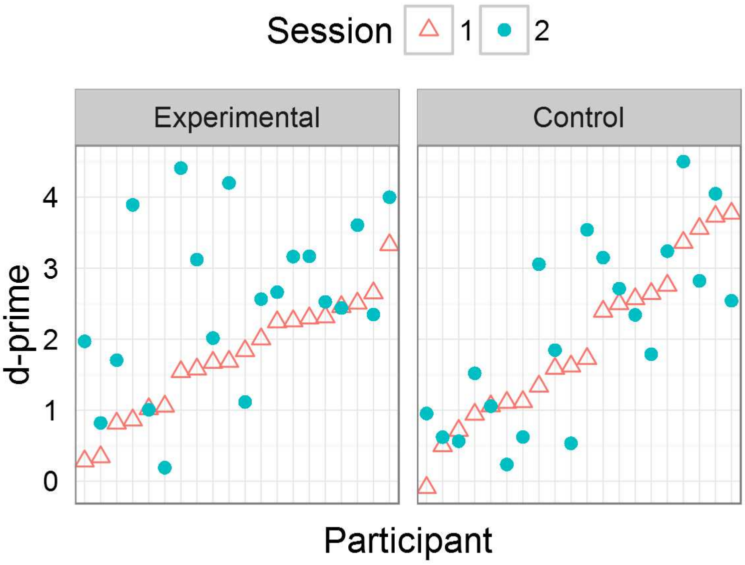

Figure 3-9. Metacognitive sensitivity for individual participants in both sessions.

ANOVA conducted on the measure of criterion revealed a close-to-significant effect of session (*F*[1, 38] = 3.36, p = .075, η_G^2 = 0.025, η_p^2 = .081), but no significant effect of group (*F*[1, 38] = 2.64, p = .113, η_G^2 = 0.047, η_p^2 = .065) or interaction of these two factors (*F*[1, 38] = 0.202, p = .656, η_G^2 = 0.0015, η_p^2 = .005). The metacognitive training did not significantly change the response bias for either the experimental (95% CI [-0.09;0.22]; *t*[19] = 0.92, *p* = .370; Hedges' *g* = 0.26, 95% CI [-0.31;0.85]) or the control group (95% CI [-0.02;0.25]; *t*[19] = 1.74, *p* = .099, Hedges' *g* = 0.33, 95% CI [-0.06;0.75]). The mean difference in the criterion change between the two groups (*M* = -0.04, 95% CI = [-0.24;0.16]) was analyzed with Welch's *t*-test and was found not significant, *t*(37.31) = -0.45, *p* = .656, Hedges' *g* = -0.14, 95% CI [-0.76;0.48]. The observed data are not surprising under the assumption of a null hypothesis that the training did not influence the criterion. See Figure 3-10.

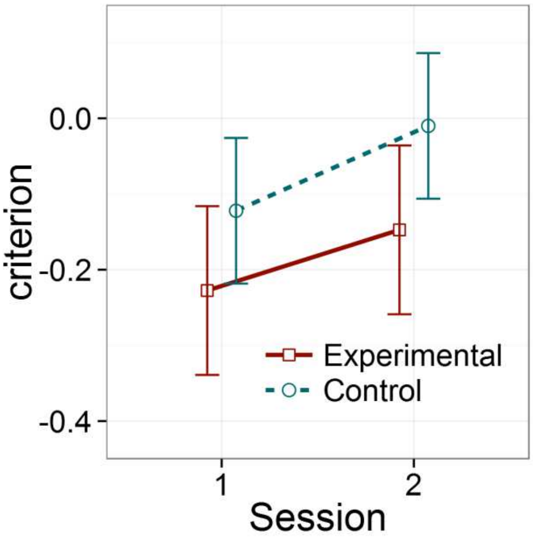

Figure 3-10. Mean criterion (bias) per session and group. *Error bars depict 95% CI for the means, corrected for dependence in measurements (Morey, 2008).*

We examined the introspection times (response times to the Decided? question), as a potential indicator of the certainty or ease of introspection. We first performed ANOVAs on the mean introspection times, with factors “group” and “session”, separately for the possible outcome classes in the Decision trials (decided-go, decided-nogo, failed-to-decide-go, but not the failed-to-decide-nogo, as there were no data in this class for some participants). In all of the three classes there was a significant main effect of session (all F[1, 38] > 11.5, p < .002), with introspection times being on average lower in the second session (M = 650 ms) than in the first session (M = 475 ms). Only in the decided-go class there was a significant interaction of group and session (F[1, 38] = 4.94, p = .032), with the decrease being bigger (95% CI = [11;238] ms) in the experimental group (M = 281 ms) than in the control group (M = 157 ms); Welch’s t-test: t[36.3] = 2.22, p = .0325. It was thus easier for participants to perform the introspective judgment after some practice with the task, and the metacognitive training improved the ease of making these judgments for trials with a reported decision to act.

We were also interested in seeing whether there would be a difference in the introspection times for the trials that were assigned to the Early and to the Late classes. For this purpose, the introspection times were analyzed as weighted averages, with the previously computed probability of a given trial belonging to the Early or the Late class serving as the weights. We then performed an ANOVA on these weighted averages with the factors “group”, “session”, “trial class” (early or late), and “reply” (“Yes”: decided-go or “No”: failed-to-decide-go). We again found a significant main effect of session (F[1, 38] = 23.2, p < .001), with introspection times being lower in the second session.

Of a larger theoretical interest was a main effect of trial class (F[1, 38] = 6.60, p = .0142), with introspection times of 641 ms for the early trials and 751 ms for the late class, 95% CI for the difference [24; 197] ms. There was also a significant interaction of trial class with reply to the Decided? question (F[1, 38] = 24.1, p < .001). When the reply to the question was "no" (failed-to-decide-go), introspection times were significantly shorter after early trials ("correct rejections", M = 568 ms) than after late trials ("misses", M = 894 ms); 95% CI = [175;477] ms, t(39) = 4.36, p < .001. The situation was reversed for the reply "yes" (decided-go), where introspection times were longer (95% CI = [16; 194] ms) for early trials ("false alarms", M = 713 ms) than for late trials ("hits", M = 608); t(39) = 2.40, p = .0216. This means that our participants came to their introspective judgments relatively faster for the correct judgments (hits and correct rejections) than for incorrect ones (false alarms and misses), suggesting that these incorrect judgments resulted rather from insufficient and uncertain introspective data than from hasty inattentive "jumping to conclusions".

3.4.3 Questionnaire data

3.4.3.1 Reports on performance

Participants reported medium difficulty of the whole experiment (median = 4, mode = 3), some amount of waiting for color change (median = 3.5, mode = 3), mid-low difficulty of the decision-making (median = 3, mode = 2), very good understanding of instructions (median = 6, mode = 7) and high overall confidence in their responses to the Decided? questions (median = 5, mode = 6).

We were interested mainly whether there is a correspondence between participants' perception of their performance and the actual performance, i.e., their metacognitive abilities. The rating of how sure people felt about their answers to the Decision? question (i.e., their confidence report on their metacognitive ability to discriminate actual decisions from falsely attributed ones) correlated strongly with the d-prime measure in session one one (r_s = .538, 95% CI [.272;.728], S = 4922, p < .001)[17] and even more so in session two (r_s = .649, 95% CI [.423;.799], S = 3738, p < .001), presumably because it was a more recent experience, Figure 3-11. This shows that even though participants did not have always good access to their mental processes (first-order metacognition), they were generally aware of that very fact to some degree (second-order, or meta-metacognition).

[17] Confidence intervals are computed for Spearman correlation with Fisher transform, which is equal to a CI for Pearson correlation on ranked data. S-value is the test statistic as computed by the cor.test function in the R statistical software. All correlation p-values are reported without correction for multiple comparisons since this is an exploratory analysis and the reader should thus interpret statistical significance with discretion.

The correlation was similarly strong and significant for both groups. Moreover, the rating of difficulty of making proper decisions was also negatively correlated with d-prime value the in session two (r_s = -.453, 95% CI [-.670;-.164], S = 15486, p = .003), but not session one (r_s = -.249, 95% CI [-.520;.068], S = 13311, p = .122).

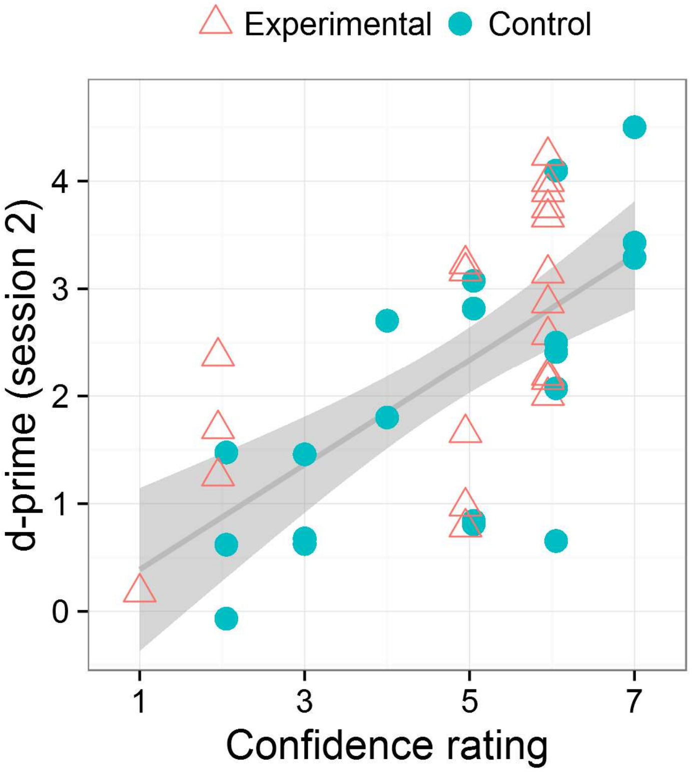

Figure 3-11. Actual metacognitive sensitivity (d') in session 2 versus confidence in metacognitive reports, "meta-metacognition". *Line depicts a linear model of the data, shaded area 95% CI.*

The participants had little insight into the fact of postponing their response to wait for the color signals, as evidenced by weak correlations of the waiting questionnaire item with several measures in the second session: the ratio of not responding within a time limit for primary responses (r_s = .205, 95% CI [-.113;.486], S = 8470.6, p = .204), with the stop-signal delay (r_s = -.116, 95% CI [-.413;.203], S = 11900, p = .475), or decide-signal delay (r_s = .055, 95% CI [-.261;.361], S = 10070, p = .735). On the other hand, participants' ratings of difficulty of the primary response condition was significantly negatively correlated with the letter discrimination accuracy in both session one (r_s = -.509, 95% CI [-.708;-.235], S = 16085, p < .001) and session two (r_s = -.509, 95% CI [-.708;-.235], S = 16087, p < .001), suggesting they were aware of this aspect of their performance to some degree.

We also found a strikingly strong and consistent inverse relationship between the ratings of feeling of control over one's actions (labeled 1: "I didn't feel to have any power over my responses, they just happened", 7: "I had a strong sense of being in control of my actions") and the ratings of difficulty of trials (which was a subsequent question, labeled 1: "Totally easy", 7: "Totally difficult") for the three different conditions (PR, Stop, and Decide). The inverse relation was present over all (40 participants × 3 conditions) data points (r_s = -.576, 95% CI [-.684;-.442], S = 453790, p < .001) and was visible in most individual participants, as evidenced by the estimated line slopes for each participant (negative for 34 participants, positive for four, zero for one, and vertical for another one; median slope -0.875 corresponding to angle of 138.8° or -41.2°, mean circular angle = 146.3° or -33.7°), see Figure 3-12.

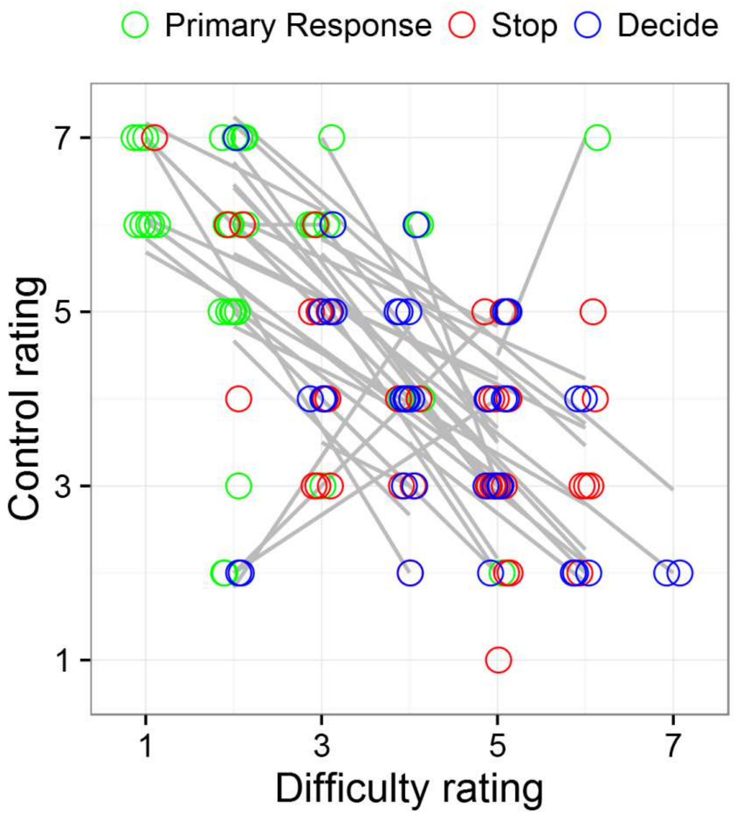

Figure 3-12. Individual ratings of difficulty and control over responses in the three conditions. *Each line represents a best fit of the three individual ratings of one participant. Data points are jittered for visualization purposes.*

The feeling of control ratings in the PR condition (median = 6, mode = 6) were higher than in the Stop condition (median = 4, mode = 3; Wilcoxon signed rank test with continuity correction, V = 589.5, p < .001) and, surprisingly, higher than in the Decide condition (median = 4, mode = 4; V =

652, $p < .001$), while there was no significant difference between the Stop and Decide conditions ($V = 258$, $p = .917$). Similarly, the difficulty ratings in the PR condition (median = 2, mode = 2) were higher than in the Stop condition (median = 5, mode = 5; $V = 57$, $p < .001$) and higher than in the Decide condition (median = 4, mode = 5; $V = 54$, $p < .001$), while there was no significant difference between the Stop and Decide conditions ($V = 234$, $p = .568$).

There was a weak relationship between the feeling of control in the Decide condition with d-prime in session 2; $r_s = .333$, 95% CI [.023;.584], $S = 7115$, $p = .036$. There were no significant differences between the two groups of participants in any of the above mentioned ratings (Wilcoxon rank sum test, all $p > .12$).

3.4.3.2 Reports on beliefs and attitudes

We performed an exploratory analysis of the scales on mindfulness, beliefs about free will and determinism, and locus of control, with relation to the main behavioral measures. However, we did not find any relationship of theoretical interest. Mainly, we failed to find a relationship between the mindful attention awareness scale (MAAS) and d-prime (session one: r[38] = .130, $p = .425$, session two: r[38] = .032, $p = .846$), suggesting that the MAAS scale might not be reflective of the introspective abilities that were probed in our study. There were no significant differences between the two groups of participants in the scales (Welche's t-test, all $p > .14$).

With respect to relationships between the individual subscales, the FAD-Plus subscales correlations to the LOC subscales exhibited correlations similar to those found previously (Paulhus & Carey, 2011), with the exception of the FAD-Plus unpredictability subscale being much more strongly correlated with the LOC chance subscale, $r(29) = .710$, $p < .001$. The MAAS scale was correlated with the LOC powerful others subscale, $r(29) = -.395$, $p = .028$.

3.5 Discussion

The aim of the present study was to test in a modified stop-signal task whether people can improve their ability to accurately assess what determinants led to their behavior: Whether their action reflexively followed after stimulus discrimination was performed or whether the action followed after a more complex processing, involving awareness of a decision signal, inhibition of a reflexive response, deliberation, and re-initiation of the action. Such ability can be considered metacognitive, because it requires a judgment about internal processing occurring between the presentation of the stimulus and the response. We replicated findings of the study on which this one was based (Kühn & Brass, 2009) that a large proportion of participants often were not able to accurately distinguish automatic from deliberated actions. People may explain (rationalize) their behavior using a priori theory that if they did something they probably decided to do it, in line with phenomena like the self-serving bias (Leary, 2007) or general desire for control (Leotti, Iyengar, & Ochsner, 2010). Nisbett and Wilson (Nisbett & Wilson, 1977) speculated that adequate feedback could disconfirm such a priori theories, potentially allowing people to make better judgments about the determinants of their behavior. The method which we adapted allowed us to provide such feedback on participants' metacognitive judgments with reasonable certainty. We found that participants performing the feedback training improved their metacognitive sensitivity overall significantly more than control participants receiving random feedback, while the feedback did not significantly change the response bias or other aspects of performance in the task (see Supplementary Results and Analyses). We thus show that metacognition, at least in some forms, is not a fixed trait but rather a plastic ability.

Nevertheless, the improvement in metacognitive sensitivity was far from perfect for most participants. Therefore although we have shown that people can possess and even improve insight into the determinants of own behavior, the long-standing skepticism about the possibility and reliability of introspective access to higher-order mental processes remains justified to some degree. It would be for instance interesting to test whether external observers could make judgments of similar accuracy as the participants themselves, as some of the more radical skeptics have suggested (Bem, 1967; Ryle, 2000). One could even argue that our study did not really improve metacognition if what the training accomplished was that participants afterwards based their judgments on their reaction times, i.e., on "public" information (Kornell, 2014). To this we reply that learning to recognize that one's inferences about the deliberativeness of an action can be wrong (disconfirmation of a priori causal theory) and that other cues such as the length of the interval between the stimulus and the response can be utilized in the inference about one's processes does count as a case of metacognitive improvement, the improvement of the knowledge of how own cognition and behavior work. In addition to this potential criticism that the training did not concern metacognition, there is a more general possible criticism, that the whole task does not involve metacognition, or introspection, which is a concept of similar breadth. Ericsson and Simon (1980) have argued that results from studies like those of Nisbett and Wilson or the present one are not surprising, because there is nothing to introspect, since we are "requesting information that was never directly heeded, thus forcing subjects to infer rather than remember their mental processes" from short term memory (p. 215). More recently, it has been argued that studies on choice blindness show that "people cannot simply rely on introspection to determine why they choose to act the way they do" but rather use inferential processes (Lind et al., 2014, p. 1202). We in principle agree with such arguments but do not consider them refuting the role of metacognition or introspection in answering questions about the determinants of one's behavior in general and in providing judgments in our study in particular. It can be argued that all judgments and perhaps even beliefs and percepts (Helmholtz, 1867) are the results of (conscious or nonconscious) inferences. However, the cues on which such inferences are based can be both internal (private) and external (public). In the present study, there certainly were internal cues present in short term memory which the participants could use to arrive at their judgments, such as a memory of perceiving the decision signal before inhibiting an automatic response, of some deliberation (e.g. thinking about the ratio of decisions up to now), and of re-initiation of the response. Participants were arguably using these cues at least to some degree, because almost all of them had above-chance sensitivity even before being subjected to the reaction-time-based training. We believe that a good conceptual framework for metacognition and introspection is needed to properly settle the presented questions.

We investigated also other metacognitive abilities via post-experiment questionnaire. Participants had little insight into the fact of postponing their responses and waiting for a stop or decision signal. On the other hand, participants were able to assess their performance in the letter discrimination task rather well. Most importantly and somewhat surprisingly, participants' confidence in their metacognitive judgments was correlated with their actual metacognitive sensitivity in the second session (Spearman's rho of .649). That means, even those participants with low metacognitive sensitivity were able to accurately assess this sensitivity, suggesting that most people had a rather good "meta-metacognition": they knew well how well they knew their own mind. Inaccuracies in this second-order metacognition tended to go in the direction of overconfidence. The correlation was similar in both groups and therefore unrelated to the training. We thus show that people have varying levels of insight into various processes and abilities.

The questionnaires also revealed an interesting finding relevant for the field studying the so-called sense of agency (Haggard & Chambon, 2012; Synofzik et al., 2013), in which studies frequently ask participants to rate their feeling of control (or judgments of authorship, etc.; many terms are being used for similar concepts) on a numeric scale, assuming that explicit ratings are valid with respect to the constructs under investigation (Valerian Chambon et al., 2013; Sato & Yasuda, 2005). It is generally assumed that we feel greater control for freely-chosen than instructed actions (Wenke et al., 2010) and when able to choose from more action options (Barlas & Obhi, 2013). In the present study we found that the ratings of feeling of control were almost equivalent to the inverse of ratings of perceived difficulty of the task in our three conditions, regardless of the actual control the participants possessed. Although the actual control in the reflexive primary response condition comes close to the characterization of our anchor label for control rating of 1 ("I didn't feel to have any power over my responses, they just happened."), the median rating was actually 6. The decision condition, in which participants were able to freely choose between two action options, received significantly lower control ratings, similar to the stop condition. We suggest that what likely occurs here is that when asked a difficult question about an unusual concept (ambiguous, not frequent in everyday interactions), people "translate" (even without being aware of that) the question as being about something more readily available, more saliently varying between the rated conditions, but still in some way plausible as an answer, in this case as being about the perceived difficulty of the task. Studies by Janet Metcalfe and colleagues(e.g., Metcalfe et al., 2013) have similarly found a strong dependence of control ratings on task performance ratings. Importantly, different people may translate the question in different ways, as demonstrated by several participants who exhibit the opposite pattern of ratings. Daniel Kahneman has called this phenomenon "attribute substitution" (Kahneman & Frederick, 2002) and we suggest that researchers should be aware of such possibility when designing and interpreting experiments.

The present study can be contrasted with other studies on metacognition using the signal detection framework, which commonly focus on reported confidence in one's performance in perceptual or memory tasks in comparison to actual performance (Fleming & Lau, 2014). Beyond the difference in topic (focus on action), our study differs in important methodological aspects. While metacognitive sensitivity is typically computed in relation to a "type 1" performance (Maniscalco & Lau, 2012), the primary reports in our study are already metacognitive in nature, and we can apply the standard "type 1" signal detection analysis to them. In addition to these reports we can also ask participants for their confidence in the reports, which we therefore call meta-metacognition. This difference is essential for the feedback training. In the confidence-rating studies an experimenter can provide a feedback about accuracy in the objective task and the participant can thus know if he or she can be more or less confident in own performance, but the experimenter cannot provide feedback about the accuracy of the metacognitive reports (in this case, whether the participant was confident or not). In contrast to that, we were able to provide feedback directly on the metacognitive reports, "reading the mind" of the participants. Although we used publicly available information (reaction times) for the "mind-reading", future studies may use private information derived for instance from neuroimaging pattern classification methods (Haynes

& Rees, 2006). Since such mind-reading is likely to remain probabilistic (due to imperfection of classification methods and because the engagement of cognitive processes can come in degrees), the approach of fuzzy signal detection theory can be useful for the study of metacognition in the future.

We want to note important limitations of the presented study. With regard to the metacognitive training, although the effect seems to be robust in the experimental group, our sample size of 40 participants allowed us to estimate the specific effect over the control group with power of only ca. 0.54, due to the time demands on participants of up to 4 hours, and the effect size interval is thus rather wide, therefore we cannot be sure about the true size of the metacognitive training effect. Further, we cannot make any claims about the duration of the effect and the possibility of transfer of the training to other domains of metacognition. Our study merely shows that such effects are possible, in the domain of metacognition of motives for behavior. With regard to the presence of frequent misattributions of deliberation to automatic behavior, we do not interpret them as telling us something significant about our free will (Kühn & Brass, 2009). The experimental situation was artificially designed to produce such misattributions and does not bear ecological validity with respect to everyday life, in which we make choices with real moral significance, compared to mere button pressing. If anything, our results weaken the “automaticity juggernaut” thesis, the claim that we run on automatic most of the time without awareness of various influences on our behavior (Kihlstrom, 2008). We show that although we created almost perfect conditions for automatic acting and post-hoc rationalization, participants were still able to provide non-confabulated reports to a large degree and were able to increase this degree via the training.

In conclusion, we designed a method allowing us to provide participants with feedback on the likely determinants of their behavior in a situation in which misattributions about the determinants are frequent. Participants often judged their actions as being the result of deliberation based on awareness of a signal to perform a decision, although reaction times strongly suggested that the action was in fact purely reactive. We show that feedback training can improve metacognitive sensitivity in such a situation without affecting response bias or other aspects of performance. We also show that participants possessed a rather good meta-metacognition, an insight into their metacognitive sensitivity, regardless of the level of this sensitivity. Contrary to skeptics, people do have some insight into their higher-order processes and causes of their actions in particular, and are able to improve such metacognitive abilities.

3.6 Acknowledgements

The research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), individual grant (WY-122/1-1) awarded to AW and the Graduate School of Systemic Neurosciences LMU. The authors would like to thank Peter Hacker, Simone Kühn, Hermann J. Müller, Dragan Rangelov, Davide Rigoni, Stephan Sellmaier, and others for fruitful discussions on the topic of this paper.

3.7 Supplementary information

3.7.1 Modelling the reaction time distributions

We modelled the decided-go RT data over all participants and both sessions of the Stop/Decide-signal task using the four candidate distributions (normal, lognormal, bimodal normal, and bimodal lognormal). The unimodal distributions were rejected based on the 1000-times repeated random data division (into a fit and a test subset), MLE model fitting, and KS tests. The mean test values and corresponding p-values for the distributions we considered are as follows: normal distribution: $D_{1225} = 0.1639$, $p < .001$; lognormal distribution: $D_{1225} = 0.1137$, $p < .001$; bimodal normal: $D_{1225} = 0.0345$, $p = .106$; bimodal lognormal: $D_{1225} = 0.0311$, $p = .183$. Additionally, the bimodal lognormal distribution has the lowest Akaike information criterion (AIC) value, which puts a cost on the number of model parameters and thus prevents selection of overly complex models (normal: AIC = 36772, lognormal: AIC = 36099, bimodal normal: AIC = 35378, bimodal lognormal: AIC = 35357). This shows that the data can be appropriately modelled as a mixture of two lognormal distributions, indicating that the decided-go trials are a result of two processes, which finish around two different times after the onset of the stimuli. The selected bimodal lognormal distribution has the following parameters, where LN refers to the probability density function of the log-normal distribution:

$$0.547 \times LN(M_1 = 6.290, SD_1 = 0.212) + (1 - 0.547) \times LN(M_2 = 7.147, SD_2 = 0.241)$$

The first component of the distribution has a median of 539 ms, the second component 1270 ms.

We performed a similar analysis for the primary response (PR) trials, considering only unimodal distributions (normal and log-normal), based on visual inspection of the data, which are clearly unimodal. The KS tests showed that the PR data can be well modelled by a unimodal log-normal distribution (the highest p-value = .0012), and that the log-normal model has a lower AIC value (409194) than the normal model (414318).

3.7.2 Descriptive statistics of SDT measures

Table 3-3. Means (and standard deviations) of SDT measures.

	Experimental group		Control group	
Measure	Session 1	Session 2	Session 1	Session 2
d-prime	1.737 (0.810)	2.546 (1.185)	1.945 (1.141)	2.085 (1.291)
criterion	-0.229 (0.267)	-0.160 (0.237)	-0.124 (0.335)	-0.012 (0.314)

3.7.3 Baseline equivalence of the groups

Even though our experiment uses a two-sessions design, thereby including a baseline which should control for potential differences between the two groups of participants, the apparent difference in decided-go RTs between the two groups in Session 1 (see Table 3-2) could nevertheless hint at a possible participant sampling bias. The difference of group means however reveals itself to be driven mostly by the high variance in this particular condition, is not statistically significant (Welch's *t*[32.7] = 1.13, *p* = .265), and is rather small between group medians (678 vs. 663 ms). We also conducted mixed-effects ANOVAs on the RTs with a within-subject factor Session and between-subject factor Group, separately for the four trial types (PR, failed-to-stop, failed-to-decide-go, decided-go). In all four cases the main effect of Group was not significant (all F[1, 38] < 0.11, p > .746), which again speaks against sampling bias. There was a main effect of Session for primary response trials (F[1, 38] = 8.34, p = .006), with RTs being on average 13 ms shorter in the second session (95% CI [4.0; 22.4]), and for the decided-go trials (F[1, 38] = 7.55, p = .009), with RTs being on average 151 ms longer in the second session (95 % CI [-266; -37]). The increase in RT in the decided-go trials was significant for the Experimental group (t[19] = 3.19, p = .005, 95% CI [-405; -84]), but not for the Control group (t[19] = 0.73, p = .471, 95% CI [-224; 108]). The interaction of Group and Session was not significant for any of the trial types (for decided-go: F[1, 38] = 2.86, p = .099; for the other three trial types: all F [1, 38] < 0.63, p > .432).

3.7.4 Equivalence in decision making behavior

It has been pointed out that the assessment of metacognitive (“type 2”) sensitivity using the signal detection framework can be biased by a change in the “primary” (or, “type 1”) performance (Maniscalco & Lau, 2012). In our case, the computation of the proposed meta-d’ measure is conceptually not applicable, because the participants’ metacognitive judgment about having decided is not contingent on another (“primary”) factual discrimination or detection judgment as is the case in e.g. perceptual confidence tasks. I.e., there is no external fact about whether the participant’s decision should be to press or not to press the button on any given trial and thus no accuracy of this decision to evaluate. Rather, the metacognitive judgment is by itself in a way the “primary” (“type 1”) factual judgment here. (The confidence rating on the certainty of these metacognitive judgments collected in a questionnaire at the end of the experiment can be considered as a “secondary” judgment.) Nevertheless, we wanted to make sure that the decision making behavior remained constant throughout the experiment, or more specifically, that any possible change between the two sessions was not different across the two groups as a result of the training. We evaluated the ratio of “Yes” and “No” responses to the Decided? question, as a factor of session and group. There was no significant effect of group (F[1, 38] = 0.86, p = .36) or session (F[1, 38] = 2.04, p = .161) and most importantly no significant interaction (F[1, 38] = 0.36, p = .554). We also evaluated the ratio of decisions for pressing and not pressing the button. Again, there was no significant main effect of group (F[1, 38] = 0.23, p = .633) or session (F[1, 38] = 0.22, p = .642) and no significant interaction (F[1, 38] = 0.018, p = .895). Additionally, we evaluated the response times to the blue “decide” signal in general (regardless whether these responses would be judged to be decisions or not). There was no significant main effect of group (F[1, 38] = 0.17, p = .687), but a significant effect of session (F[1, 38] = 8.97, p = .0048), with response times being slower in the second session (M = 656 ms, SD = 122 ms) than in the first session (M = 613 ms, SD = 103 ms). However, there was no significant interaction (F[1, 38] = 0.83, p = .369). These results suggest that the behavior related to the decision making was not significantly influenced by the metacognitive training.

3.7.5 Predicting responses on decision trials

We were interested whether we would be able to predict responses of participants on Decision trials from their behavioral performance preceding the response. That is, we wanted to know on what factors depends (1) whether participants would press a button or withhold their response (go or nogo, “responded” variable, coded 1 and 0, respectively), (2) whether they would answer to the Decided? question in affirmative or negative (decided or failed-to-decide, “decided” variable, coded 1 and 0, respectively), and (3) in particular, whether they would falsely attribute early responses as decisions or label them correctly as non-decisions. We constructed one logistic regression model for each of these cases. For the “responded” variable, we chose the following predictors: mean (exponentiated) and standard deviation of individual log-transformed reaction times for the PR trials, ratio of go vs. no-go from beginning of a session up to the current trial, ratio of decided vs. failed-to-decide up to the current trial, presence of response on the last decision trial, answer to the Decided? question on the last decision trial, condition of the preceding trial (PR, Stop, or Decide), what key would be used for a correct letter discrimination response on the current trial, and signal delay (between presentation of a black stimulus and change of color to blue) on the current trial. For the “decided” variable we added also the presence or absence on the current trial, response time on the current trial, and response time for the Decided? question. To quantify the predictive ability of the models we performed a 10-fold cross-validation, and computed mean squared errors, mean classification accuracy (in what fraction of trials the actual response is correctly predicted), mean classification sensitivity index (d’), and mean classification bias (criterion).

3.7.5.1 Predicting the presence of response

The statistics of predictors for the “responded” variable are in Table 3-4. The mean squared error of the model was 0.178, classification accuracy 0.752, but in close inspection, sensitivity index was low (0.660) while bias was high (-1.864), meaning the model was almost always right when the response was present, but almost always wrong when the response was absent. Influence of most of the predictors is not surprising, such as that of the Responded ratio and Responded previous (participants were instructed to decide for responding or not responding roughly equally often; influence of reported decision on last decision trial can be related to that) or the signal delay (longer delay means higher chance of inhibiting a response). The influence of whether the correct response key is left or right for the letter discrimination task is less explicable and could suggest a small handedness bias. When response was provided, the previous decision trial was reported as decision more often than when response was not provided (54.0 vs 45.6 %), suggesting that participants tended to inhibit their response more often if they failed to inhibit it previously.

Table 3-4. Predictor variables for the presence of a response.

Predictor	Coefficient	Std. Error	z-value	p-value

(Intercept)	-4.401	0.437	-10.068	< .001 ***
PR RT Mean [s]	0.296	0.553	0.534	0.593
PR RT SD [log ms]	1.140	0.644	1.772	0.076
Responded ratio	6.988	0.352	19.868	< .001 ***
Decided ratio	0.407	0.378	1.074	0.283
Responded previous	-0.197	0.072	-2.737	0.006 **
Decided previous	0.355	0.061	5.781	< .001 ***
Preceding condition: PR	0.081	0.139	0.580	0.562
Preceding condition: Stop	0.121	0.146	0.826	0.409
Correct response: Right	-0.129	0.050	-2.570	0.010 *
Signal delay [s]	-1.252	0.338	-3.707	< .001 ***

3.7.5.2 Predicting the report of decision

The statistics of predictors for the "decided" variable are in Table 3-5. The mean squared error of the model was 0.104, classification accuracy 0.856, sensitivity index was 2.253 and bias was 0.316, meaning it was possible to predict the answer to the Decided? question from the preceding behavior to a moderate degree. Again, the influence of some of the predictors is not surprising, like the presence of response (absence is almost always reported as decision), RT for the response (longer RTs suggest deliberation), or the signal delay (longer delay increases chance of inhibition and deliberation). It is noteworthy that reporting a decision was more common after previously reporting an absence of decision compared to presence (on average 53.7 vs. 50.4 %), but that reporting a decision occurred when the ratio of decisions so far was higher than when reporting an absence of decision (53.7 vs. 50.7 %). Also reporting a decision was more common when no response was given on the preceding decision trial (which was likely labeled as decision) compared to when it was given (53.7 vs. 51.4 %), and similarly, the ratio of responses on decision trials so far was lower when decision was reported than when absence of decision was reported (74.4 vs. 75.9 %). It is interesting that decisions were reported more often when the correct response key for the letter discrimination task was the same as for reporting presence of decision, suggesting a response bias, possibly in cases of uncertainty, but this effect was rather small (52.2 vs. 51.8 %). There was also a small relationship with the time needed to answer the Decided? question, which was on average 13 ms longer before answering "no". It is important to note that interpreting the predictors in isolation is difficult, because their predictive role can become evident only after accounting for (regressing out) other predictors in the model. This is certainly the case of the individual mean RT for PR trials, which is a significant predictor, but not by itself, as it can probably account for some individual differences between participants; similarly for the SD of PR RTs.

Table 3-5. Predictor variables for the report of decision.

Predictor	Coefficient	Std. Error	z-value	p-value
(Intercept)	3.020	0.569	5.308	< .001 ***
PR RT Mean [s]	-11.863	0.784	-15.130	< .001 ***
PR RT SD [log ms]	1.980	0.852	2.323	0.020 *
Responded ratio	3.251	0.435	7.480	< .001 ***
Decided ratio	7.392	0.551	13.418	< .001 ***
Responded previous	-0.411	0.091	-4.526	< .001 ***
Decided previous	-0.439	0.079	-5.580	< .001 ***
Preceding condition: PR	-0.113	0.179	-0.629	0.530
Preceding condition: Stop	-0.177	0.189	-0.934	0.350
Correct response: Right	-0.217	0.066	-3.308	< .001 ***
Signal delay [s]	3.987	0.421	9.473	< .001 ***
Responded current	-9.027	0.203	-44.525	< .001 ***
RT current [s]	7.298	0.237	30.830	< .001 ***
RT Decided? question [s]	0.161	0.063	2.559	0.011 *

3.7.5.3 Predicting the report of decision for early responses

Finally, we created a model for early response trials in the decision condition, specifically those trials for which there was only 1% chance of belonging to the late distribution. That is, we used a crisp threshold which was supposed to capture trials which were in fact reflexive response, and we were interested whether it could be predicted based on the above-mentioned factors, whether these trials would be classified as real decisions (i.e., performing a false alarm) or as non-decisions (correct rejection). This analysis could allow us to see if there are some factors which lead participants to confabulate or to perform accurate introspection. Surprisingly, the prediction factors for only these early decision trials were similar to the previous model for all decision trials, see Table 3-6. Of interest was the response time for the Decided? question, with faster responses for correct rejections (on average 553 ms) than for false alarms (724 ms), suggesting that participants were less certain in falsely attributed cases. Also, participants reported decision more often when the same key as was used for the “Yes” answer was also used just before that in response to the letter task (24.2 vs. 20.0 %), showing that even such an irrelevant contextual factor was able to influence introspective judgments. Decision was also reported slightly more often when decision was also reported on the preceding decision trial (22.7 vs 21.5 %) and less often when response was provided on the preceding decision trial (21.5 vs 24.0 %). However, the overall predictive and classification power of the model was rather low, with mean squared error of the model 0.145, classification accuracy 0.790, sensitivity index 0.881 and bias 1.281.

Table 3-6. Predictor variables for the report of decision for early responses.

Predictor	Coefficient	Std. Error	z-value	p-value
(Intercept)	-6.211	0.618	-10.049	< .001 ***
PR RT Mean [s]	-11.539	0.945	-12.207	< .001 ***
PR RT SD [log ms]	2.011	0.961	2.093	0.036 *
Responded ratio	2.911	0.488	5.961	< .001 ***
Decided ratio	6.722	0.599	11.225	< .001 ***
Responded previous	-0.483	0.101	-4.781	< .001 ***
Decided previous	-0.424	0.089	-4.769	< .001 ***
Preceding condition: PR	-0.010	0.199	-0.049	0.961
Preceding condition: Stop	-0.101	0.210	-0.483	0.629
Correct response: Right	-0.340	0.074	-4.621	< .001 ***
Signal delay [s]	3.813	0.465	8.206	< .001 ***
RT current [s]	8.290	0.470	17.628	< .001 ***
RT Decided? question [s]	0.453	0.069	6.587	< .001 ***

3.7.6 Pilot Experiment

The pilot experiment was very similar to the current one, but there were slight differences in the design and it was considerably shorter. We tested 30 volunteers. The experiment consisted of five blocks with 96 trials in each block, that is, 480 trials per participant, taking about 40 minutes. Compared to our current design, there was a fixed time limit of 2000 ms in the primary response condition instead of the individual time limit, and also there were no feedback messages warning about possible waiting strategies. The results showed a bimodal distribution of decided-go reaction times, see Figure 3-13. Maximum likelihood estimation and Kolmogorov-Smirnov testing determined a similar bimodal log-normal model as in the current study, only with a more prominent second peak, indicating fewer misattributed deliberations, which could be attributed to lower response time pressure in this pilot study:

$$0.292 \times LN(M_1 = 6.408, SD_1 = 0.216) + (1 - 0.292) \times LN(M_2 = 7.202, SD_2 = 0.239)$$

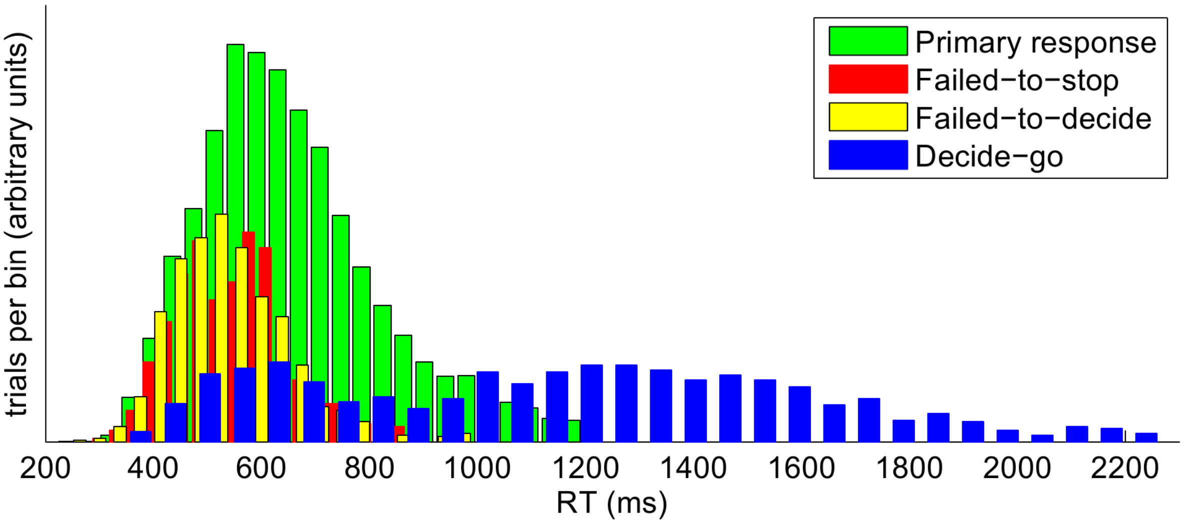

Figure 3-13. Histogram of reaction times over all participants in a pilot experiment.

We calculated two crisp classification thresholds to separate the two log-normal distributions such that there would be only a 1% chance of a trial to belong to the other distribution, using a similar method as in the original study this one was based on (Kühn & Brass, 2009). The "early" threshold was at 770 ms and the "late" threshold at 1004 ms. These thresholds were then expressed as quantiles of the aggregate primary response distribution, such that the early threshold was at quantile 0.6049 (z-score 0.2661) and the late threshold at quantile 0.9627 (z-score 1.7829). These two quantiles were employed as thresholds for the feedback messages during the metacognitive training in our current study. Note that the early threshold was relatively conservative with respect to classifying a response as false attribution of decision, being only somewhat longer than the mean of the individual's PR distribution, and that there was a relatively large "middle" interval of uncertain classification in between these two thresholds. We employed this method because we needed to give crisp feedback messages on each trial and we wanted to provide these messages only when we were relatively highly certain about the classification. The mean PR RTs in the first session of our experiment were shorter than in the pilot because of the prevention of waiting strategies, producing thresholds of lower RT values, i.e., classifying more trials as being decisions. Importantly, the method that was used to classify these trials and compute our main dependent measures was independent of the method used to determine the training thresholds. The results of this pilot experiment also showed rather large individual differences in the amounts of falsely attributed deliberation to reflexive responses, see Figure 3-14.

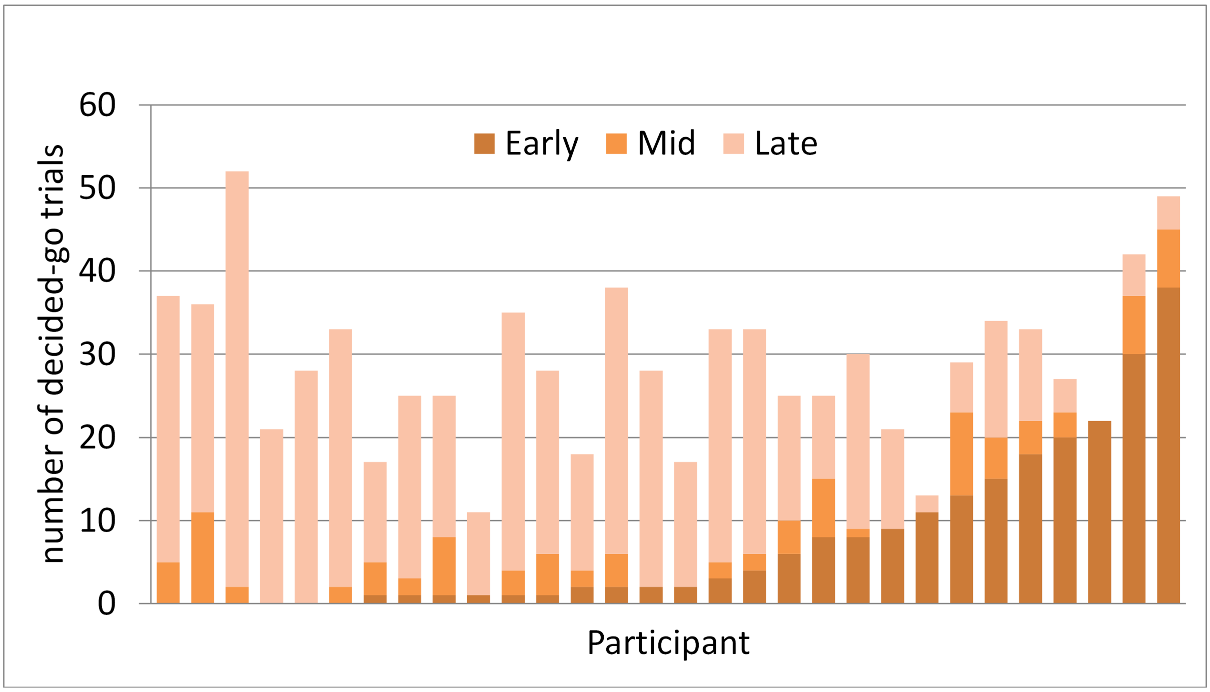

Figure 3-14. Individual differences in the numbers of false or correct attributions of decision.

3.7.7 Participant feedback questionnaire

1a	How difficult was this task for you? (1 = Totally easy, 7 = Totally hard)
1b	Was it easier or harder now compared to the first session? (1 = It was much easier now, 7 = It was much harder now)
2	Are you aware of waiting for the change of the color? (1 = I didn't wait at all, 7 = I waited always)
3	Did you have any strategies? If yes, what strategies? (Like preparing the response for the next blue trial in advance, or alternating between two responses..)
4a	Was it hard for you to make proper decisions in the blue trials? (1 = It was totally simple in all the cases, 7 = It was really very hard for me most of the time)
4b	Was it easier or harder now compared to the first session? (1 = Much easier, 7 = Much harder)
5	Did you have a clear understanding of the instruction "to decide" in the blue trials? (1 = I have no idea what kind of decision I was supposed to make and/or I don't have a clear understanding of the concept of "decision" in general. 7 = It was totally clear to me what it means to make a decision between two options and I perfectly knew what kind of mental operation I was supposed to do.)
5b	Was it more or less clear now compared to the first session? (1 = Much clearer, 7 = Much less clear)

6	Were you sure about your responses to the “Decided?” question? (1 = I had no idea whether I really decided most of the time, 7 = I was totally sure in all the cases whether I really decided about pressing the button.)
6b	If you remember, were you now surer about the responses to the “Decided?” question than in the previous session? (1 = I was much surer now, 7 = I was much less sure now)
7a	How much did you feel in control over your responses to different types of trials? (1 = I didn’t feel to have any power over my responses, they just happened. 7 = I had a strong sense of being in control of my actions.) Normal trials:
7b	Red trials:
7c	Blue trials:
8a	How difficult were for you the different types of trials? (1 = Totally easy, 7 = Totally difficult) Normal trials:
8b	Red trials:
8c	Blue trials:
9	Did you pay attention to the messages you were receiving after the blue trials? What did you think about them? Were they correct? Do you think they made you change your behavior in the experiment?

4 Study 2: Expect to Be Distracted

Expect to be distracted: Prediction of salient distractor by action and cue attenuates its interference

Ondřej Havlíček[1,2], Hermann J. Müller[1,3], Agnieszka Wykowska[1,4]

[1] Department of Psychology, General and Experimental Psychology Unit, Ludwig-Maximilians-Universität, Munich, Germany

[2] Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität, Munich, Germany

[3] Department of Psychological Sciences, Birkbeck College, University of London, London, UK

[4] Engineering Psychology, Division of Human Work Sciences, Luleå University of Technology, Luleå, Sweden

Author contributions: OH, AW, HJM conceived and designed the study. OH performed the research and analyzed data. OH, AW, HJM wrote the paper.

4.1 Abstract

Prediction and attention are crucial factors influencing how the human brain processes sensory inputs. While there is a vast amount of evidence showing benefits of predicting task-relevant sensory information, it is less clear whether and how predicting distracting irrelevant sensory events affects performance. We investigated how task-irrelevant distracting events that are predicted either by one's own action or by an external cue influence attentional processing in a visual search paradigm with an additional salient singleton (distractor). We found that the interference related to the salient task-irrelevant singleton (attentional capture) can be attenuated when the singleton is predicted, independent of the type of prediction. This pattern of results can be explained better within the predictive coding framework than the optimal motor control theory or by the idea that prediction mandatorily allocates attention to task-irrelevant items (the "attentional white bear" hypothesis).

Keywords: attentional capture, attentional white bear, forward model, motor control, predictive coding, sensory attenuation, visual attention

4.2 Introduction

It is well established that attention can enhance the processing of certain aspects of visual information, such as particular features or locations in the visual field. While the underlying mechanisms would usually improve everyday performance by concentrating processing on behaviorally relevant information, performance may also be harmed when salient information is selected, or 'captures attention', that is actually task-irrelevant (Hickey, McDonald, & Theeuwes, 2006; Theeuwes, 1992; Theeuwes, Atchley, & Kramer, 2000; Yantis, 1993). There has been a considerable debate as to whether and how such bottom-up-driven processes can be overcome (Bacon & Egeth, 1994; Eimer & Kiss, 2008; Müller, Geyer, Zehetleitner, & Krummenacher, 2009; Müller, Reimann, & Krummenacher, 2003; Theeuwes, 2010; Wykowska & Schubö, 2010, 2011). One possible way for top-down processes to modulate bottom-up mechanisms may be based on predictive information regarding aspects of the task and stimuli presented. Although there has recently been a coming-together of the fields of attention research and research on the effects of prediction on perception (Clark, 2013; Feldman & Friston, 2010; Hohwy, 2012; Jiang, Summerfield, & Egner, 2013; Kok, Rahnev, Jehee, Lau, & de Lange, 2012; Summerfield & Egner, 2009), how predictive mechanisms influence attentional selection remains unclear. On the one hand, attention is typically thought to enhance information processing, e.g., at a specific location in the visual field (Posner, 1980), by mechanisms of sensory gain control (Hillyard, Vogel, & Luck, 1998). On the other hand, predicted stimuli have been reported to produce reduced neural responses compared to unexpected ones (Alink, Schwiedrzik, Kohler, Singer, & Muckli, 2010; Summerfield & Egner, 2009; Todorovic, van Ede, Maris, & de Lange, 2011). That is, the notion of attentional enhancement of sensory signals would appear to conflict with that of prediction-related attenuation (Bubic, von Cramon, & Schubotz, 2010; Spratling, 2008; Summerfield & Egner, 2009). There have been attempts in the field to reconcile this apparent conflict; for example, Kok and colleagues (Kok et al., 2012) showed that item predictability can interact with top-down attentional orienting towards task-relevant items, with neuronal response in visual cortex being increased for predicted, as compared to non-predicted, relevant targets and decreased for predicted irrelevant items. However, a comprehensive theory reconciling attention and prediction effects is still missing.

4.2.1 Attentional white bear? Predicting the irrelevant item

While there is a vast literature devoted to prediction of target locations or target-defining features, as well as the interaction between goal-directed attention and prediction, relatively little is known about the interaction between prediction and bottom-up attentional orienting towards salient but irrelevant items. Conceivably, knowing about the location or featural description of a distracting item could help ignore it; though, paradoxically, attention could also be especially drawn to this item, increasing its distracting effect. The latter effect has actually been reported in the experimental literature (Tsal & Makovski, 2006) and termed “attentional white-bear phenomenon” (AWB). In Tsal and Makovki’s study (Tsal & Makovski, 2006), participants were engaged in a flanker task, i.e., their task was to discriminate a target flanked by two distractors. In the condition of interest, the locations of the flanking distractors were kept the same throughout a block of trials. On some trials within this block, a secondary task display was presented instead of the flanker display, consisting of two simultaneously presented dots, one of which appeared at the same position at which a flanking distractor would normally be located. The task on this type of trial was to judge which one of the two simultaneously presented dots appeared first. The key finding was that participants more frequently reported that it was the dot presented at the location where a distractor stimulus was expected to appear – an observation that proved to be robust to various experimental manipulations (Lahav, Makovski, & Tsal, 2012). The authors argued that the first item selected is likely to be the distractor, in part because the very instruction to ignore the distractor will represent it, as a kind of ‘template’, in visual working memory, biasing the allocation of focal attention towards a distractor actually appearing in the display – in the same way as when trying not to think about a white bear makes one focus on its very mental image. Directing attention to the distractor, in turn, made participants perceive the dot appearing at that location as occurring first, in line with the 'prior-entry' literature (Shore, Spence, & Klein, 2001; Spence & Parise, 2010; Titchener, 1908).

The AWB hypothesis has been explored in recent years with conflicting findings. For instance, Beck and colleagues (2011) found that maintaining the distractor representation in visual working memory resulted in a higher rate of eye movements towards the distractor during serial visual search. Similarly, Olivers (2009) documented several cases of memory-driven attentional capture. In contrast, Woodman and Luck (2007) found that observers managed to not attend to such items. Furthermore, Arita and colleagues (2012) showed that participants were able to use distractor feature information to bias attention away from such non-target items (held in memory) during serial visual search. Similarly, Dhawan and colleagues (2013) reported that asking participants to hold a location forbidden for saccades in memory resulted in successful attentional inhibition of this location.

4.2.2 Different ways to induce prediction

4.2.2.1 Prediction by spatial cueing

The most straightforward way to examine the effects of distractor prediction would be to make the distractor predictable by a cue. Again, opposing hypotheses can be formulated regarding the results of such a manipulation. One might, for instance, expect that cueing the salient but task-irrelevant item would actually lead to more pronounced orienting towards it and thus increased interference, because even task-irrelevant symbolic cues can give rise to attentional orienting (Hommel, Pratt, Colzato, & Godijn, 2001). However, intuitively, one might also expect that participants can learn to use such information to orient away from the distractor, e.g., by inhibiting its location. Ruff and Driver (2006), in a Posner-type task (i.e., there were two possible stimulus locations, one on the left and one on the right of a central fixation marker), presented participants with a valid pre-cue to the target location on each trial. On some trials, a distractor would appear contralateral to the target. Informing participants that such a distractor would be present on the upcoming trial resulted in improved performance; by contrast, telling participants that a distractor would be absent yielded no benefit. From this, the authors thus argued that having foreknowledge about the distracting stimulus "specifically allowed participants to counteract the impact of the … distractor" (Ruff & Driver, 2006, p. 531). Similarly, Munneke and colleagues (2008) as well as Chao (2010) reported that pre-cueing the distractor location can lead to inhibition of its location, yielding beneficial effects on performance. In contrast, Buckolz and colleagues (2006) failed to find inhibition of distractor location via a spatial cue. Thus, taken together, the evidence regarding the effects of cueing distracting stimuli is mixed.

4.2.2.2 Prediction by action

Prediction by direct spatial cueing is not the only way to make participants expect an irrelevant item. There are several other sources of prediction or expectations, a prominent one being one's own actions (Waszak et al., 2012). Throughout our lives, we continue to learn what sensory outcomes can result from our actions. It would thus be reasonable to expect that predictability of such action effects would allow us to better guide our attention to task-relevant stimuli and away from distracting stimuli. Actually, however, little is known as yet about the specific impact of action-effect prediction on attention and the mechanisms involved.

It is thought that predictions related to actions attenuate the strength of the actions' sensory consequences (Waszak et al., 2012; Wolpert & Flanagan, 2001). A paradigmatic case in point is the finding that participants find it hard to experience the sensation of being tickled if they control some robotic arm that does the tickling, whereas they feel more tickled if a temporal delay or trajectory perturbation is introduced into the motion of the robotic arm (Blakemore, Frith, & Wolpert, 1999). Sensory attenuation has also been demonstrated in the auditory domain (Baess, Horváth, Jacobsen, & Schröger, 2011; Hughes et al., 2013; Weiss, Herwig, & Schütz-Bosbach, 2011). Regarding the visual domain, though, the experimental evidence is scarce, with, to our knowledge, only one study reporting a decrease in sensitivity for self-produced stimuli (action-related prediction), compared to stimuli predicted by auditory tones (accompanied by a non-predictive action) (Cardoso-Leite et al., 2010).

There are two prominent groups of theories that make predictions regarding signal attenuation for action-produced stimuli. On the one hand, optimal control theory (Blakemore et al., 1999; Wolpert & Flanagan, 2001) posits that a stimulus, such as a salient but task-irrelevant distractor, predicted by a forward model should be perceived as being less strong, and thus produce less interference compared to when it is not predicted or when it is predicted only by external events (e.g., a cue) that are presumed not to enter the forward model. By contrast, another group of theories, associated with the notion of predictive coding, postulates a more general predictive mechanism (H. Brown et al., 2013; Friston, 2011; Pickering & Clark, 2014; Van Doorn, Hohwy, & Symmons, 2014), based on which one would not expect a difference between the various sources of predictive information given that they predict the same outcome with the same ‘precision’ (i.e., the inverse variance of outcomes over multiple trials). According to these theories, predicted input can be both up- and down-regulated sensorily, depending on its expected precision (Hohwy, 2012; Kok et al., 2012). However, it remains an open question what expected precision should be assigned to a salient irrelevant distractor. Precision weighting is thought to be (at least partially) open to task demands (e.g., generating an expectation of precision for some spatial region or a perceptual feature) and thus conceptualized in terms of endogenous attention (Feldman & Friston, 2010). Task-irrelevant stimuli should therefore have reduced expected precision (Kok et al., 2012). On the other hand, the distractor is also physically salient – and, as such, thought to be processed as a high-precision stimulus, given the (innate or acquired) prior expectation that strong stimuli have a high signal-to-noise ratio and are thus more precise (Feldman & Friston, 2010).

4.2.3 Aim of study

On this background, the present study was designed to address two related questions: First, does general predictability of a highly salient but task-irrelevant visual stimulus increase or decrease the interference it generates in a situation of efficient visual search, that is, when both the target and the distractor pop out and, thus, strongly compete for selection? This question is related to the debate of whether predicting irrelevant stimuli does enhance or attenuate their processing. Second, would the effects of prediction differ when the distractor is predicted by one’s own action, as compared to when the information as to the presence and location of the distractor is provided by an external cue?

4.2.4 Design

While prediction of the presence and location of an item by a directional (spatial) cue can be implemented in a straightforward manner, sensorimotor contingencies between an action and a stimulus are arbitrary and have, thus, to be learned through (experiencing the) repeated coupling of the action with the stimulus. To create the most reliable association, we employed just two possible actions, using the middle finger of the left or, respectively, the right hand, and coupled these with just two possible sensory outcomes, namely: ‘production’ of the distractor at a particular one of two possible locations in the search display. In the actual task after initial learning, an additional (‘neutral’) action: pressing a central button with the index fingers of both hands at the same time, was used to produce displays without the distractor. In order to avoid interference with the learned associations, we adopted the same response procedure as implemented by Cardoso-Leite and colleagues (2010): two possible response options were presented alternately on the screen (one at a time) until the participant selected one of them using the neutral action. Given that this response procedure does not allow for speeded reactions, only performance accuracy, rather than reaction time (RT), was available as dependent measure. That is: distractor interference was measured in terms of the difference in accuracy on distractor-absent versus distractor-present trials.

Note though that, in the literature, interference by salient singletons has most reliably been observed in terms of increased RTs (Theeuwes, 1992; Yantis, 1993). Nevertheless, assuming that the RT cost generated by the distractor originates from the process of visual selection (rather than from, e.g., response selection), presenting the display only briefly and masking it at the end of its presentation would make it less likely for the target to be processed if attention had first been captured by the distractor in a probabilistic manner (as a result of the higher salience of the distractor compared to the target; (Zehetleitner, Koch, Goschy, & Müller, 2013)). To our knowledge, there are only a few studies measuring distractor interference in briefly presented displays using accuracy. For example, Kiss and colleagues (2012) reported increased error rates due to distractor presence in displays presented for 200 ms, but no mask was used in their paradigm. Gibson and Jiang (1998) used a similar paradigm where search displays were presented only for 86 ms and masked, but failed to find a cost in accuracy – their only dependent measure. However, in their study, this is likely attributable to the fact that their search task was very effortful, performed in a serial fashion.

Given that demonstrating distractor interference with short presentation times has turned out to be difficult in the past, we created conditions under which distractor interference is strong in general, thus making it more likely to be ‘reflected’ in a measure of performance accuracy. Concretely, we employed a so-called ‘compound’ visual search task (Duncan, 1985) in which the target-defining and the response-defining stimulus properties are separated. In the present task, the target-defining feature was odd-one-out shape, that is, participants were asked to locate the singleton shape (=target) in the display, and the response was to be made with respect to the orientation of a line probe inside the target. The non-target items were all homogeneous, square-shaped. Regarding the shape of the target, any of the target’s four corners could be missing; that is, it was not always precisely the same shape, which was expected to increase participants’ reliance on a ‘singleton search’ strategy (Bacon & Egeth, 1994). That is, participants, rather than looking for a specific pre-defined shape, would search for any odd-one-out shape. Searching for a singleton target has been reported to increase the interference from task-irrelevant singleton distractors (Bacon & Egeth, 1994; Lamy & Egeth, 2003). In the present study, the irrelevant singleton was made more salient than the target by virtue of its increased luminance. We used dense displays where both target and distractor would be surrounded by neutral items from all sides to further increase the salience of the singleton distractor (Rangelov, Müller, & Zehetleitner, 2013). Display durations were determined by a staircase procedure run prior to the experiment proper, individually for each participant, such that target discrimination accuracy was at a set threshold level (of 71% correct).

Our aim was to examine the change in the influence of the distractor on target selection when the distractor was predicted by either a cue or an action, as compared to a baseline condition without any predictive information. These three conditions were presented blocked, in separate sessions, with the action prediction condition being the last one. The reason for this was that the action condition had to be preceded by a sensorimotor contingency learning period, like in other studies implementing an initial acquisition phase followed by a test phase (Herwig & Waszak, 2012; Richters & Eskew, 2009). As the association thus learnt (in the acquisition phase) could have influenced any trials that would come after the action prediction session, the baseline and cue prediction conditions could not be presented after the action prediction condition (doing otherwise would have led to a confounding of the results).

Hughes and colleagues (2013) suggested that many studies investigating the influence of action-related prediction on perception might be confounded, because the participants' action did not just predict the identity of the resulting stimulus (specific configuration and properties of items); rather, it also allowed for temporal prediction as to when the stimuli would appear and even for temporal control over the (onset of the) stimuli. To address these methodological concerns, the same three actions (left, right, and neutral action) that were used in the action prediction condition were also employed, in the same ratio (1:1:2), in the baseline and cue prediction conditions to start the trials. Actions in the two latter conditions were completely unrelated to the 'identity' of (i.e., the presence and location of distractors in) the subsequent displays, and, beyond making all conditions equal in these respects, they served the purpose of helping participants learn to randomly produce the three actions in the specified ratio for the action prediction condition, in which the actual ratio of actions was important. To control for effects of exposure to the distractor stimuli (e.g., (Müller et al., 2009)) during the association learning period, we included simple distractor observation periods prior to the baseline and the cue prediction periods.

Taken together, we implemented a novel paradigm designed to measure the interference effects of a salient singleton distractor using accuracy and, at the same time, allowing for prediction of the distractor by a to-be-performed action, in addition to a condition with an endogenous cue. Our first question, namely: whether predictability of the distractor would influence the interference with target selection, can be answered by examining whether any kind of predictive information would either increase or decrease the accuracy cost of the distractor in the baseline condition. The answer to the second question, whether there is a difference between using an external, endogenous cue and a to-be-performed action as the source of the predictive information, would be provided by a difference in the accuracy cost between the two types of prediction.

4.3 Materials and methods

4.3.1 Participants

Because the task proved to be rather difficult, data collection was ongoing until we had ‘usable’ data from 30 participants, where ‘usable’ was defined a-priori as accuracy above chance level in each combination of predictive condition and distractor presence. Overall, 44 participants were tested, but 14 failed to meet the accuracy criterion. Participants were randomly assigned to one of two action-effect contingency mapping groups, that is: *natural mapping*: left key press producing a distractor on the left side (and right key a distractor on the right side), versus *inverse mapping*: left key producing a distractor on the right side (or vice versa), with 15 participants in each group.

The number of participants was based on a priori considerations. Since one possible hypothesis is a zero difference between prediction by cue and prediction by action, in order to minimize the possibility of a false negative finding, it was necessary to estimate the expected effect size that the action prediction should have on top of the cue prediction according to theories invoking action-effect prediction via forward models. To our knowledge, there is only one study comparable in its aim and design (Cardoso-Leite et al., 2010), which reported a significant sensitivity reduction when the stimuli were predicted by action (d_z=0.793) but not when they were predicted by a tone (d_z=0.247). Although both effects were investigated in two different groups of participants (of unequal size), the additional effect of action prediction on top of cue prediction can be estimated as the difference between these two effects (d_z=0.546). To detect an effect of such size with a reasonable power of at least .8 (J. Cohen, 1988) in a within-subject design, we would a-priori need to test 29 participants. So as to be able to counterbalance the mapping of two possible actions and their effects, and thus to have two groups of participants of equal size, we decided to test 30 participants.

Participants' age range was 19-34 (M = 24.6) years; all of them were right-handed; and nine were male. All participants self-reported normal or corrected-to-normal vision. They were paid € 8 per hour or opted to receive a course credit. The experiments were conducted at the Experimental Psychology laboratory of the LMU Munich. All experimental procedures consisted of purely behavioral data collection of healthy adult participants, without involving invasive or potentially dangerous methods. The study was approved by the ethics committee of the LMU Psychology Department, in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Data were stored and analyzed anonymously. All participants provided written, informed consent.

4.3.2 Apparatus

Participants were seated in a dimly lit and sound-attenuated room, in front of a CRT monitor (LaCie Electron 21/108, refresh rate 100 Hz, resolution 1024×768 pixels), at a viewing distance of 58 cm, with their heads stabilized using a chin rest. A standard keyboard was used to collect responses. Participants were instructed to use their left middle finger to press the C key (left response key), the right middle finger to press M key (right response key), and to press the spacebar always using both index fingers at the same time for a neutral response.

4.3.3 Stimuli

The E-Prime software (Psychology Software Tools Inc., Sharpsburg, PA, version 2.0 Professional) was used to set up and present the stimuli. The search display consisted of 20 grey square items (size 1.05°×1.05° of visual angle, luminance 13.3 cd/m^2, RGB [64, 64, 64]) against a black background (luminance 1.24 cd/m^2); the items were positioned around three (imaginary) concentric circles (equally spaced, outer diameter 11.7°), with a grey fixation cross in the center. In the visual search displays, one of the items: the target had one of the four corners cut off. On some trials, a bright grey square – that is: a distractor – was present (luminance 58.4 cd/m^2, RGB [160, 160, 164]). Additionally, each of the items contained a probe: a black line (size 0.6×0.1°) oriented either vertically or horizontally. The target and distractor were limited to locations on the middle circle (diameter 7.2°). Pattern masks were presented to mask each individual search display item (at the end of display exposure). The masks consisted of a black-line cross and a diamond inside a square (as if there were both a vertical and a horizontal probe line inside the item and all corners of the item were cut off), and was of the same color and luminance as the distractor. See Figure 1 for a depiction.

4.3.4 Procedure

The experiment consisted, in the main, of the three blocked conditions: baseline, cue prediction, and action prediction, in which participants performed a variation of essentially the same task; this task will be described first, followed by the specific differences among the conditions and other details.

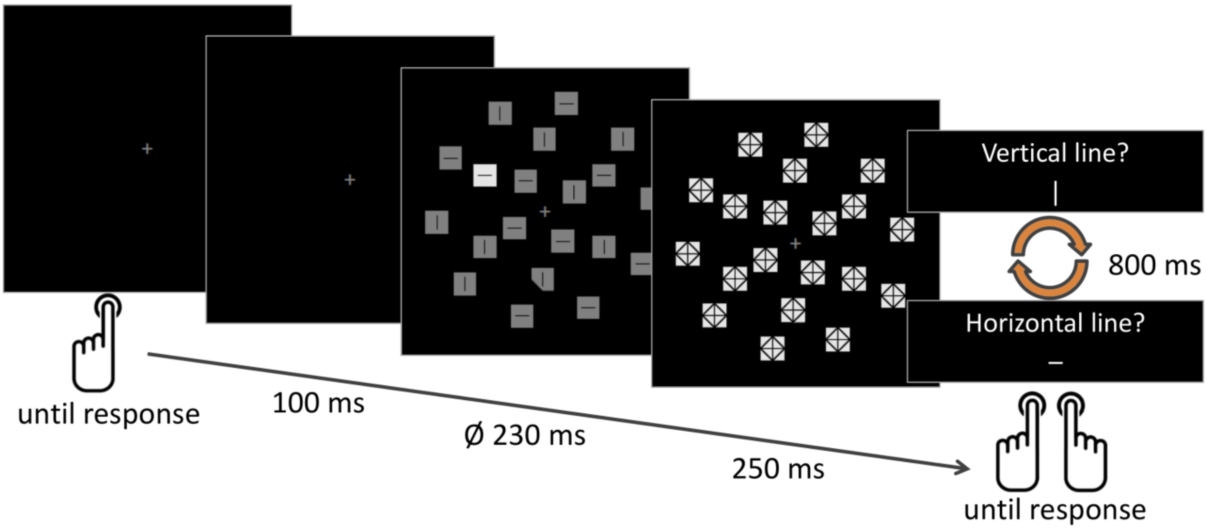

Figure 4-1. Basic trial sequence. *Each trial begins with a fixation cross (or a cue in the cue prediction condition) displayed until the press of a left, right, or a neutral key. Afterwards, a search screen is presented for a duration previously determined by a staircase and subsequently masked. Participants select the correct target probe orientation from two alternating options using a neutral key.*

4.3.4.1 General visual search task

The basic paradigm employed in the three blocked conditions was an additional singleton compound visual search task. Each trial began with a grey fixation cross displayed in the middle of the screen. Participants could then, at any time, press a key – which, after a delay of 100 ms, produced the search display on the trial (Figure 4-1). Each item in the display contained a line probe oriented randomly in either vertical or horizontal direction. The instruction was to search for a shape singleton: a square with a random corner cut off, and report the orientation of the line inside this target. The target could appear at one of the six locations on the middle circle, twice as likely at the top and bottom locations, relative to the lateral positions. (This specific ratio was chosen to allow for a comparison with a planned ERP study, which would require such a ratio of midline and lateral target occurrences.) In one half of the trials, a luminance distractor was randomly displayed at either the top-left or bottom-right location. The display was presented only for a brief period of time, determined individually in a pre-experimental staircase procedure (M = 227 ms, SD = 83). The display was then masked for 250 ms. Next, messages “Horizontal line?“ and “Vertical line?“ with a picture of the respective line orientation were presented alternately on the screen (800 ms / message), repeating until the participant made a response, that is, pressed the neutral key using both hands at the time when the answer deemed correct was displayed. Feedback was provided in the case of an incorrect response (in the form of a red “minus” sign presented for 1000 ms). Afterwards, a blank screen was displayed for an inter-trial interval (ITI) of 250-550 ms (uniform random distribution).

Participants were asked to press one of three different keys to initiate each trial: the left, the right, or the neutral key. They were instructed to choose among the keys at will, but to press the neutral key about twice as often as the other keys, optimally in a ratio of 25% : 25% : 50%. In the baseline and cue prediction conditions, the identity of the key had no effect on the task. In the *cue prediction* condition, the fixation cross at the start of a trial was replaced by a central symbolic cue, which could be: a left arrow (“<”) sign, indicating that the distractor would be displayed at the top-left location; a right arrow (“>”) sign, indicating a distractor at the bottom-right location; or a minus (“-”) sign, indicator that no distractor would be presented. The cue was displayed until a participant initiated a trial with a button press; the cue was invariably (100%) valid. Participants were explicitly informed about this and told to use the information provided by the cue to help them to perform the task better. In the *action prediction* condition, the key used to start the trial determined the presence and location of the distractor. The neutral key produced no distractor, while the left and right keys would produce the distractor at one of the two usual (i.e., the top-left or bottom-right) locations. This action-effect contingency was counterbalanced across participants (between-participants factor “contingency group”: natural mapping vs. inverse mapping): for one half of the participants, the left key would produce the distractor at the top-left position and for the other half at the bottom-right position; and conversely for the right key. Participants were explicitly informed about this action-effect contingency.

There were six blocks of trials in each of the three conditions, with each block consisting of 32 trials, yielding a total of 192 trials per condition. After each block, participants were given a feedback about their key press ratio and allowed to rest for a while.

4.3.4.2 Association task

The action prediction condition was preceded by an association phase, intended for participants to learn the sensorimotor contingencies between an action (button press) and the observed effect (display with a distractor) prior to performing the action prediction condition proper. The task in the association phase was to randomly press the left or the right key in a period of time during which an empty screen with a fixation cross was displayed, in a (key press) ratio of 50% : 50% approximately, at a pace of about one press per two seconds. The key press produced a display, after a delay of 100 ms, similar to the search display, with the exception that there were no probes inside the items and no target. A distractor was always present at one of the two locations according to the participant's contingency group. The duration of this display was 600 ms. In one out of eight trials, the central fixation cross in the search display had the color red: a catch trial. On such trials, participants were required to immediately press the neutral key with both their index fingers at the same time. The idea behind this was to make sure that participants payed attention to the displays. The duration of the display as well as the response window for the catch trials was 1000 ms. In case of an incorrect response or a failure to respond, a red "minus" sign would appear for 1000 ms.

There were seven blocks of trials, each block consisting of 64 trials, that is, 448 association trials in total. The number of association trials was chosen based on the Cardoso-Leite et al. (2010) study. After each block, participants were given feedback about their key press ratio and allowed to rest.

4.3.4.3 Exposure task

To control for possible effects of exposure to the stimuli, a session similar to the association session was administered before both the baseline and the cue-prediction visual search condition. In these sessions, participants did not use a key to produce the displays but were simply observing a stream of displays. First, a fixation cross appeared for 600 ms. This was followed by a search display of the same kind as in the association session, for 600 ms. The distractor was randomly located at one of the two usual locations on each trial. One eighth of the trials were again catch trials, in which the fixation cross was of red color and participants had to press the neutral key as fast as possible. The duration of the display as well as the response window for the catch stimulus was 1000 ms. In case of an incorrect response or a failure to respond, a red "minus" sign would appear for 1000 ms.

There were six blocks of trials, each block consisting of 64 trials, that is, 384 exposure trials in total. After each block, participants were allowed to rest.

4.3.4.4 Staircase

Before the actual experiment, the search display durations were determined individually for each participant. An adaptive staircase procedure was used to find the individual thresholds. The visual search task described above was used; however, only the neutral key was used to start the trials and a distractor was always present, located randomly at any of the six locations on the middle circle. The search display duration started at a set value of 400 ms and was increased by one step size in case of an error and decreased by one step size in case of two successive correct responses. This staircase rule aims at approximately 71% accuracy threshold. The step size was 80 ms until the 4th reversal point (error after a correct response or vice versa), 40 ms until the 6th reversal, and then kept at 10 ms. The procedure terminated after 16 reversals, and the final display duration was calculated as the average duration across the last 10 reversal points, rounded to a multiple of ten.

4.3.4.5 Overall structure of the experiment

Participants began with the staircase phase in order to establish the display duration to be used in a subsequent practice phase. This practice phase had the same structure as the actual experiment, but was limited to two blocks of eight trials per each of the six experimental phases. After practice, participants performed the staircase procedure once more, and the value obtained was introduced in the actual experiment. The order of the experimental parts that followed was: exposure, visual search (baseline), exposure, visual search (cue prediction), association, visual search (action prediction). After the experiment, a one-question ‘questionnaire’ was administered asking participants how specifically they had used the information provided by the cue. The whole experiment took between 1.5 and 2 hours to complete, including instructions and all breaks.

4.3.5 Analysis

The staircase procedure was used to adjust the difficulty of the visual search task for each participant. However, it was still possible that the staircase failed to find the correct threshold setting, making the task too difficult for some participants, possibly resulting in (near-) chance level performance. To verify that participants were actually able to perform the main task above chance level, individual performance was assessed using a binomial test for each combination of prediction type condition and distractor presence. If the accuracy in any of these combinations was not significantly higher than what could be expected by chance ($\alpha = .05$), the data of this participant were excluded from analysis. Additionally, several trials in the action prediction phase had to be excluded due to technical issues (error in the program) during data acquisition. However, this affected only 2.57% of trials, on average, in this particular condition.

We tested our hypotheses using a 2 × 3 repeated-measures analysis of variance (ANOVA) on mean accuracies, with the factors "prediction type condition" (baseline, cue prediction, action prediction) and "distractor presence" (distractor absent, distractor present), followed up with individual two-tailed paired-samples t-tests comparing the cost of distractor presence on accuracy between prediction type conditions. Of most interest to our first main question – whether the distractor would exert a lesser or greater influence when predicted – was the difference in distractor interference between the baseline and each of the two prediction type conditions. Our second question – that is, whether the type of prediction influences the magnitude of distractor interference – was examined by analyzing the difference in distractor interference between the two prediction type conditions. Distractor interference was quantified as the difference in accuracy between distractor-absent and distractor-present trials, and can be conceptualized as the incidence of attentional-capture events.

4.4 Results

4.4.1 Baseline distractor interference effect

Before examining what the effect of prediction on distractor interference was, we tested whether the irrelevant singleton did actually interfere with search performance in the baseline condition of our paradigm. The answer was affirmative: accuracy was significantly higher in the absence of the distractor (M = .812, SD = 0.074) than in its presence (M = .755, SD = 0.071); $t(29) = 4.449$, $p < 0.001$, 95% CI [0.031, 0.083], $d_z = 0.812$.

4.4.2 The effect of distractor prediction on attentional capture

Next, an ANOVA was performed on all conditions. This analysis revealed a significant main effect of distractor presence, $F(1, 29) = 21.0$, $p < .001$, $\eta_G^2 = .048$, $\eta_p^2 = .420$. Participants displayed generally lower accuracy in the presence of the distractor (M = .780, SD = 0.075) compared to its absence (M = .814, SD = 0.077). Follow-on t-tests revealed that the interfering effect of the distractor was significant not only in the baseline condition, but also in the cue prediction condition (*t*[29] = 2.743, p = .010, 95% CI [0.0065, 0.045], d_z = 0.501), while being marginal in the action prediction condition (*t*[29] = 1.85, p = .074, 95% CI [-0.0019, 0.038], d_z = 0.339), see Table 4-1. Furthermore, accuracy was overall significantly different across the different prediction type conditions, *F*(2, 58) = 3.676, p = .031, η_G^2 = .016, η_p^2 = .112, see Table 4-2. Importantly for the purpose of the present study, and of relevance for our hypothesis that predictability of the distractor would modulate its impact on task performance, the interaction between distractor presence and prediction type condition was significant, *F*(2, 58) = 4.509, p = .015, η_G^2 = .012, η_p^2 = .135. See Figure 4-2.

Table 4-1. Descriptive statistics for the accuracy for all levels of predictability condition and distractor presence.

Prediction type condition	Distractor absent			Distractor present		
	M	SD	95 CI	M	SD	95 CI
Baseline	0.8122	0.0743	0.7844 - 0.8399	0.7552	0.0710	0.7287 - 0.7817
Cue	0.8146	0.0700	0.7884 - 0.8407	0.7889	0.0713	0.7623 - 0.8155
Action	0.8147	0.0882	0.7817 - 0.8476	0.7964	0.0789	0.7669 - 0.8259

M = mean, SD = standard deviation, 95 CI = 95% confidence interval for the mean, uncorrected.

Table 4-2. Descriptive statistics for the accuracy for all levels of predictability condition.

Prediction type condition	M	SD	95 CI
Baseline	0.7837	0.0776	0.7636 - 0.8037
Cue	0.8017	0.0713	0.7833 - 0.8201
Action	0.8055	0.0835	0.7840 - 0.8271

M = mean, SD = standard deviation, 95 CI = 95% confidence interval for the mean, uncorrected.

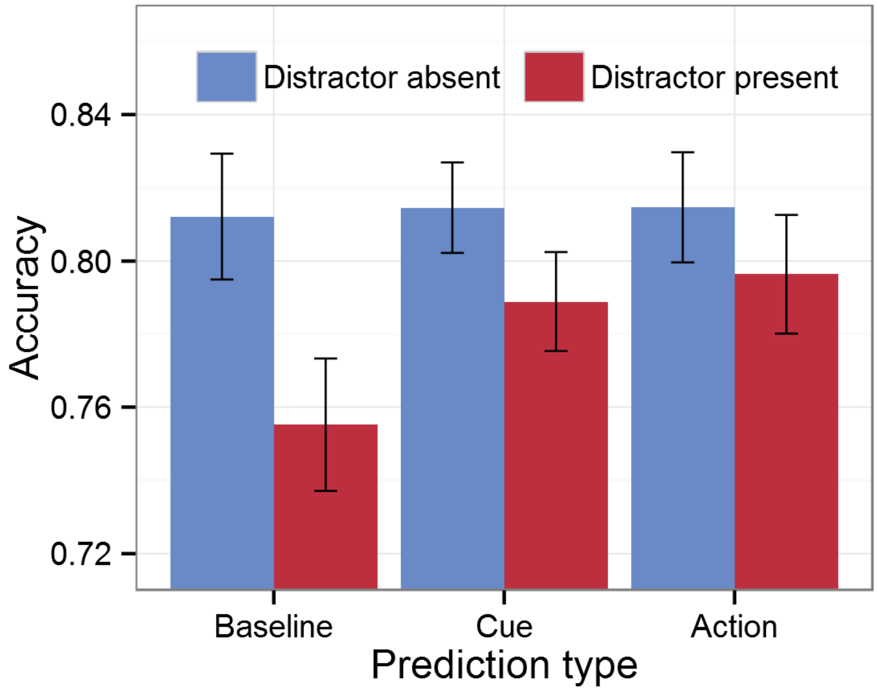

Figure 4-2. Accuracy per prediction type condition and distractor presence and absence. *Prediction of the distractor by either a cue or an action decreases the interference caused by the distractor. Error bars show 95% confidence intervals corrected for dependence in measurements (Morey, 2008).*

4.4.3 Types of prediction

To directly address whether the type of prediction had an influence on the reduction of attentional capture, the interaction was followed up with t-tests on the size of distractor interference (mean accuracy on distractor-present trials *minus* mean accuracy on distractor-absent trials). There was a significant difference in the magnitude of distractor interference between (i) the baseline and cue prediction conditions (*t*[29] = 2.07, p = .048, 95% CI [0.00032, 0.062], d_z = 0.377), with the interference being less marked in the latter condition (baseline: M = .057, SD = 0.070, vs. cue prediction: M = .026, SD = 0.051), and (ii) between the baseline versus the action prediction condition; *t*[29] = 2.71, p = .011, 95% CI [0.0095, 0.068], d_z = 0.495, again with less marked interference in the latter condition (baseline: M = .057, SD = 0.070, vs. M = .018, SD = 0.054). This pattern indicates that both kinds of predictive information were effective in attenuating the distractor's interference. However, the difference in distractor's interference between the cue and action prediction conditions was not significant, *t*[29] = 0.656, p = .517, 95% CI [-0.016, 0.031], d_z = 0.120. Note that the predictive information influenced mainly the distractor-present trials; there was no significant difference in accuracy among the prediction type conditions for distractor-absent trials (*F*[2, 58] = 0.037, p = 0.963, $\eta_G^2 < .001$, $\eta_p^2 = .001$).

4.4.4 Additional analyses

4.4.4.1 Power and Bayesian analyses

One of our main questions concerns the difference in distractor interference between prediction based on a cue versus an action. Since we did not observe a statistically significant difference between these two conditions (p = .517), we cannot make any firm conclusions as to the actual presence or absence of the effect. However, we can analyze the likelihood of having obtained a false negative finding, given that we had an a priori expectation for the effect size of $d_z = 0.546$ (Cardoso-Leite et al., 2010). First, our achieved statistical power for such an effect size is .824, which makes the chances of a false negative finding relatively small, without however eliminating such a possibility. Second, Bayes factor analysis using a non-informative prior (Cauchy with a recommended scale r = 0.707) allowing for a wider range of expected effect sizes (Rouder, Speckman, Sun, Morey, & Iverson, 2009) showed that there is 4.22-times more evidence for the null hypothesis of no effect.

4.4.4.2 Sensorimotor contingencies

To test for the possibility that the effect of the action prediction was not due to the assigned sensorimotor contingencies, we performed a mixed-design two-way ANOVA (on the action prediction condition data), with accuracy as the dependent measure, distractor presence as a within-subject factor, and contingency group (natural vs. inverse) as between-subject factor. Neither the main effect of the contingency group (*F*[1,28] = 1.407, p = .246, $\eta_G^2 = .043$, $\eta_p^2 = .048$) nor, importantly, the interaction of contingency group with distractor presence (*F*[1,28] = 0.593, p = .448, $\eta_G^2 = .0022$, $\eta_p^2 = .021$), was significant.

4.4.4.3 Learning effects

Because we presented conditions in a fixed sequence, it is possible that the improvement in accuracy in the presence versus the absence of a distractor could be explained by simple learning. This would manifest itself as an interaction between the magnitude of distractor interference and time in at least one of the conditions; specifically, interference would decrease as a function of time. To examine for such an effect, we conducted an ANOVA on interference magnitude (accuracy on distractor-absent minus distractor-present trials) with the within-subject factor prediction type condition. Time was represented by a numeric within-subject variable "block number" (there were 6 blocks of trials in each condition). While the main effect of prediction type condition was (still) significant, the main effect of block number was not (F[1,29] = 0.289, p = .595, η_G^2 = 0.0019), and there was no interaction of block number with prediction type condition (F[2,58] = 0.235, p = .791, η_G^2 = 0.0036). See Figure 4-3. This means that learning alone cannot explain the observed reduction in distractor interference in the cue and action prediction conditions.

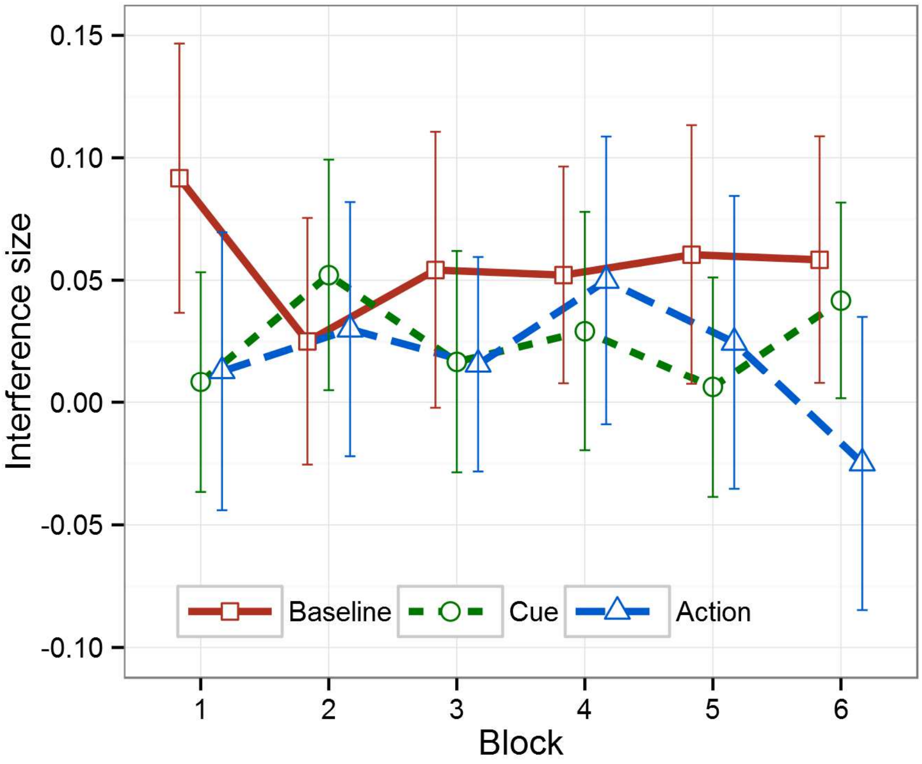

Figure 4-3. Distractor interference size as a function of time and prediction type condition. *There was no significant effect of time (discretized into six blocks of trials) on distractor interference size (accuracy for distractor absent minus distractor present trials) within any of the prediction type conditions (baseline, cue, or action). Error bars show 95% confidence intervals corrected for dependence in measurements (Morey, 2008).*

4.5 Discussion

The present study aimed at examining the interaction between prediction and attention processes, in particular: whether predicting the presence and location of an item that is task-irrelevant but attention-capturing would interfere with task performance to a larger or smaller degree, relative to when no prediction regarding the irrelevant item is possible; furthermore, whether the type of predictive information (cue- vs. action-based) would play a role for how attention is allocated to the irrelevant but salient item. To answer these questions, we employed an additional-singleton compound visual search task, with only briefly presented and subsequently masked stimuli, with the task design ensuring that distractor interference could be measured reliably in terms of accuracy costs. It could be argued that accuracy measures are not as sensitive as RT measures to detect small effects, but this may be outweighed by the fact that accuracy, in contrast to RTs, is little influenced by late, response-related processes and more directly reflects mechanisms operating at early, perceptual processing stages (of stimulus selection and discrimination) – which were of central interest in relation to the questions addressed. In this paradigm, we manipulated the way in which the presence and location of the distractor was predicted (via either an external cue or an internally-generated action), while controlling for such potential confounding factors as the presence of action, cognitive load, temporal prediction, temporal control, and exposure to the distractor stimuli across all prediction type conditions. We found that, for both prediction types, the distractor interference was reduced compared to a relatively large interference effect in a baseline condition that did not provide any information about the location or presence of a distractor. Interestingly, the interference reduction was comparable in magnitude whether the distractor was predicted by an external cue or by an action.

Predictive information of either type about the *absence* of a distractor had no effect compared to the baseline, suggesting that the prediction indeed influenced the processing of the distractor item (i.e., the performance improvement was not due to any other facilitatory processes related to the provision of predictive information as such).

4.5.1 Baseline attentional capture effect

The effect pattern obtained in the baseline condition clearly revealed the additional salient but irrelevant singleton to interfere with the search performance. This demonstrates that by strengthening the conditions for attentional capture, accuracy can also be a reliable measure of the interference cause by the distractor. In our design, conditions for attentional capture were strengthened by using a relatively salient, easily detectable shape target, thus promoting parallel search within a wide attentional window. Inducing a wide attentional window was crucial, in line with Theeuwes (2010), who argued that during serial search, the attentional window is set to ‘narrow’ – so that the distractor is then less likely to fall inside the window and capture attention.

Also, the shape orientation was different across trials, arguably providing an incentive for participants to adopt a singleton, rather than a feature, search mode (Lamy & Egeth, 2003). Hence, the present results confirm that sufficient saliency of the irrelevant singleton, inducing a singleton search mode, and a broad setting of the attentional window are critical factors for attentional-capture effects to be observed with briefly presented and masked displays and accuracy as dependent measure.

4.5.2 The effect of predicting the irrelevant singleton

Furthermore, our results showed clear reductions of the magnitude of distractor interference with both kinds, or sources, of predictive information – which cannot be attributed to a simple learning effect, as the distractor interference effect did not show a significant decrease over time within the individual blocked conditions (i.e., the interference effect remained at a similar level across blocks within both the baseline condition and the cue condition). Instead of a gradual decrease of the interference effect with time, there was a rather abrupt drop in interference magnitude between the baseline and the cue conditions (Figure 3). This pattern provides evidence against “white-bear”-type processes (at least for the present paradigm) and in favor of the idea that predicting the presence and location of an irrelevant singleton can actually attenuate attentional capture, rather than making the distractor an even stronger attractor for attention (compared to situations with no information about the distracting item).

With respect to the case of prediction by external cue, the finding of prediction-driven attenuation of distractor interference is particularly interesting. The information provided by the cue was task-irrelevant. Participants were instructed to use the cue information in any way they wanted to help them perform the task better. When asked about their strategy at the end of the experiment, most participants reported having made no use of it. Because only three participants reported consistent usage of the cue, while the cue clearly had a positive effect on performance for most participants, we assume that the cue was actually being used by the majority of participants in some automatic manner, without intention and awareness. That people can extract cue information without being aware of this has been reported previously (Decaix, Siéroff, & Bartolomeo, 2002; Peterson & Gibson, 2011). A similar case can be made for the action prediction condition, in which participants presumably lacked a reason to deliberately and consciously guide their attention according to the button they pressed (although participants were not explicitly questioned about this at the end of the experiment). Interestingly, this fits recent findings of Chisholm and Kingstone (2014): they varied the level of participants' awareness about the distractor and observed a performance benefit in the condition of intermediate emphasis, when participants were only told about the presence of distractors, as opposed to conditions in which they were either completely unaware of the distractors or explicitly asked to avoid them. This explicit instruction can be what causes the attentional "white-bear phenomenon", possibly by representing the distractor in visual working memory, which then biases perceptual selection. Note that Chisholm and Kingston used an overt-attention task (participants were required to make a saccade to the target in order to discriminate a probe inside of it), though they pointed out that it remains to be seen whether similar findings would be obtained for covert attention as well. Our data provide a tentative affirmative answer to this question.

4.5.3 Difference between the types of prediction

Hughes and colleagues (Hughes et al., 2013) proposed that there are several kinds of predictive processes involved in tasks designed to study action-effect anticipation. Our design controlled for temporal-control and temporal-prediction processes in all conditions, as participants were able to start each trial at a time of their choosing and the action-effect occurred after a constant delay. The difference in performance between the baseline and cue prediction conditions was supposed to reveal the influence of what Hughes and colleagues refer to as "identity prediction" processes, that is, predicting the stimulus (and its properties) in a general manner (not necessarily related to motor processes). And importantly, the difference between the cue and action prediction conditions was supposed to directly reflect the contribution of internal motor-prediction processes, possibly in line with forward models of action-effect prediction (Waszak et al., 2012; Wolpert & Flanagan, 2001).

We observed significantly reduced attentional capture in the action prediction condition compared to the baseline, which suggests that either the top-down influence of knowledge of distractor presence and location (i.e., identity prediction processes) or the internal motor-prediction processes, or both, had an influence on the processing of the irrelevant stimulus. However, we did not observe any significant difference between the cue and action prediction conditions. There may be several reasons for the latter, which will be discussed below.

4.5.3.1 Optimal motor-control theories

The forward-model account would predict an effect of the action on top of that of the cue (Hughes et al., 2013). That we failed to observe such an additional effect could, conceivably, be simply a false negative finding, owing to lack of statistical power. However, this is unlikely to be the case, as we achieved a power of .824 for observing the expected effect size. Additionally, the Bayes factor analysis we performed shows that there is substantially more evidence for the null-hypothesis of no effect than for the hypothesis of an effect.

Alternatively, it is possible that our design was too different from that of Cardoso-Leite et al. (2010), in that we were not providing predictive information about a near-threshold stimulus but rather about a highly salient one. Forward-model theories postulate that predicted action-effects (predicted sensory input) are subject to sensory attenuation, but the specific mechanism of this attenuation is not clear. It is, for instance, possible that the sensory signals are attenuated in a non-linear fashion, that is, there can be sensitive ranges for which the attenuation is most pronounced (near-threshold stimuli) and ceilings at which attenuation is very limited (strong stimuli). Indeed, Zehetleitner et al. (2013) showed that the probability that a distractor will capture attention on a given trial is a psychometric function of the difference in salience between the distractor and the target: if the distractor is much more salient than the target, a small decrease in distractor salience – as would be due to the presumed sensory attenuation in our study – will not translate into any, or only a very small, reduction of the probability of attentional capture. Thus, if the forward-model-driven sensory attenuation effect is too small to detect or not present at all, we need to assume that the reduced amount of interference observed in the present action condition is either due to the same mechanism that is at play in the cue prediction condition, or due to a wholly different mechanism that just happens to produce an effect of similar magnitude – neither of which is accounted for by the forward-model theory.

4.5.3.2 Predictive coding

Although we cannot exclude the possibility of multiple predictive processes being at work in our experiment, the most parsimonious explanation is that both types of prediction work in the same (or a very similar) fashion, possibly being two instances of a more general cognitive process. Since prediction appears to be present in almost all brain regions and networks, it has been argued that this may reflect a more general principle of how the brain works (Bubic et al., 2010; Friston, 2010; Pickering & Clark, 2014). This idea is built upon the framework of predictive coding, or more generally: predictive processing (Clark, 2013). According to this framework, sensory signals are the product of prediction error and expected precision, at multiple levels of the cortical hierarchy (Feldman & Friston, 2010). While the calculation of the prediction error is relatively straightforward, based on the sensory input and expectations regarding this input, its precision – the weight that will be assigned to this error – must be determined by the system, and this can be conceptualized as ‘attention’. Highly salient stimuli are a-priori expected to be precise, which corresponds to exogenous attention; on the other hand, precision can also be set according to the task demands, as is the case with endogenous attention (Hohwy, 2012). In general, this theory postulates attenuation of predicted sensory signals, because there is a smaller prediction error that needs to be explained. Moreover, it has been shown that prediction interacts with endogenous attention, as evidenced by increased BOLD signals in early visual cortices for attended predicted stimuli, which compares with reduced activation for unattended predicted stimuli (Kok et al., 2012). The same pattern could be predicted for exogenous attention as well, but this has not been tested as yet. With respect to the account of precision weighting, our design directly pitted endogenous (task-relevant target) and exogenous (salient distractor) attention against each other, while manipulating the predictability of the latter. Within this framework (Feldman & Friston, 2010), the fact that we found distractor interference to be reduced in the predictive conditions would mean that the target was better able to compete for attentional resources vis-à-vis the salient distractor, but that, nevertheless, the precision-weighted prediction error was not invariably higher for the target than for the distractor. This is so because the distractor clearly managed to capture attention on a portion of trials, as the interference effect was not completely abolished in the prediction conditions.

Most importantly for our study, the predictive-coding account would indeed predict no difference between the two sources of prediction – because both of them lead to the same prediction with the same accuracy regarding the identity of the stimuli, their location, and timing. The only difference might be in the weight that is attributed to these sources, according to their respective precision. It seems reasonable to assume that after a sufficient amount of practice, both sources will be assigned a very similar expected precision and should thus influence performance in a similar way. However, this account is far from complete, for instance because prediction error and precision have to be computed at each level of the processing hierarchy and different features are predicted and need to be considered to identify an item as a target or as a distractor – which opens up new questions for future research.

4.5.4 Conclusions

In conclusion, we introduced a novel paradigm investigating the effects of two types of prediction – action-related and cue-induced prediction – regarding the presence and location of a salient task-irrelevant object in an attentional-selection task. The presence of a distractor was found to have a negative impact on target selection, but this impact was attenuated, rather than enhanced, when the presence and location of the distractor was predicted either by an external, endogenous cue or by a to-be-performed action – which argues against the ‘attentional white-bear hypothesis’, though possibly because the irrelevant singleton was not actively maintained in working memory. Importantly, we did not find a difference between the two types of prediction. We propose that this pattern of results can be explained within the framework of predictive processing, which has recently been shown to be able to account for various aspects of attention (den Ouden, Kok, & de Lange, 2012; Feldman & Friston, 2010; Kok et al., 2012). However, we note that the results could also be in line with other explanations, including the role of forward-model-specific prediction. In sum, our study contributes to the growing picture of how prediction can improve attentional selection, even in the case of distracting, task-irrelevant stimuli, and regardless of whether the stimuli are a learnt by-product of our own actions or whether they are predictable by external environmental cues.

4.6 Acknowledgements

The research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), individual grant (WY-122/1-1) awarded to AW and the Graduate School of Systemic Neurosciences LMU. The authors would like to thank Karl J. Friston, Jakob Hohwy, and Florian Waszak for fruitful discussions on the topic of this paper.

4.7 Additional information

Here we present results that were not included in the manuscript.

4.7.1 Distance effects

A variety of studies have reported an inhibitory surround around a focus of attention (Gaspar & McDonald, 2014; Hopf et al., 2006; Tombu & Tsotsos, 2008), therefore if a distractor captures attention, the target is less likely to be processed in the close proximity of the distractor (Koch, Müller, & Zehetleitner, 2013; Mounts, 2005). The analysis of distance effects would also allow us to exclude non-spatial effect (e.g., filtering costs) as the sole mechanism of the distractor interference. We therefore divided the trials in which the distractor was present into three groups according to the distance between the target and the distractor (approximately, 3.6, 6.2 and 7.2 degrees of visual angle), see Figure 4-4. After collapsing over all three prediction type conditions we have indeed found an evidence of lower accuracy for targets at the closest distance of 3.6° (M = .771, SD = 0.0737) than for distance 6.2° (M = .792, SD = 0.0707), which was a hypothesis-based prediction and therefore tested with a one-tailed t-test; *t*[29] = -2.02, p = .0264, one-tailed, d_z = 0.369.

To test whether distractor prediction influences the effect of distance, we conducted an ANOVA on mean accuracy in distractor present trials with “distance” as a within-subjects factor and “prediction presence” as a within-subject factor (levels: “baseline” and “prediction present”, after collapsing cue prediction and action prediction type conditions together) and we have indeed found a significant interaction; *F*[2, 58] = 3.488, p = .046, η_G^2 = 0.020. The main effect of distance was not significant; *F*[2, 58] = 0.617, p = .489, η_G^2 = 0.0063. If we follow the interaction up with t-tests, we can see that prediction significantly improved accuracy for the short (*t*[29] = -3.145, p = .0038) and middle (*t*[29] = -2.561, p = .0159) distances, but not for the longest distance (*t*[29] = 0.704, p = .487), see Table 4-3. No such interaction with distance was found between the individual prediction-present conditions (levels: “cue prediction”, “action prediction”); F[1, 29] = 0.0022, p = .963, η_G^2 < 0.0001. This similarity (Figure 4-4) between the accuracies in cue and action prediction conditions further suggests that they are driven by the same process.

Alternative explanation could be that these effects are not due to distance but rather due to relative target-distractor laterality. When testing a model with a factor “laterality” (within-subjects; levels: “same hemi-field”, “midline”, “opposite hemi-field”) instead of the factor “distance” we found no significant main effect of laterality (*F*[2, 58] = 0.614, p = .545, η_G^2 = 0.006) and no interaction of laterality with prediction presence (*F*[2, 58] = 1.167, p = .318, η_G^2 = 0.005). This pattern of data suggests that these results are not exclusively due to laterality but rather due to distance.

Table 4-3. Descriptive statistics for the accuracy for all levels of predictability condition and target-distractor distance.

Prediction type condition	Distance	M	SD	95 CI
Baseline	No distractor	0.8122	0.0743	0.7844 - 0.8399
Baseline	3.6	0.7424	0.0857	0.7104 - 0.7744
Baseline	6.2	0.7639	0.0899	0.7303 - 0.7974
Baseline	7.2	0.7806	0.1585	0.7214 - 0.8398
Cue	No distractor	0.8146	0.0700	0.7884 - 0.8407
Cue	3.6	0.7833	0.0807	0.7532 - 0.8135
Cue	6.2	0.8074	0.0828	0.7765 - 0.8383
Cue	7.2	0.7556	0.1434	0.7020 - 0.8091
Action	No distractor	0.8147	0.0882	0.7817 - 0.8476
Action	3.6	0.7933	0.1012	0.7555 - 0.8311
Action	6.2	0.8055	0.0977	0.7690 - 0.8420
Action	7.2	0.7685	0.1444	0.7146 - 0.8224

Distance is in degrees of visual angle, M = mean accuracy, SD = standard deviation, 95 CI = 95% confidence interval for the mean, uncorrected.

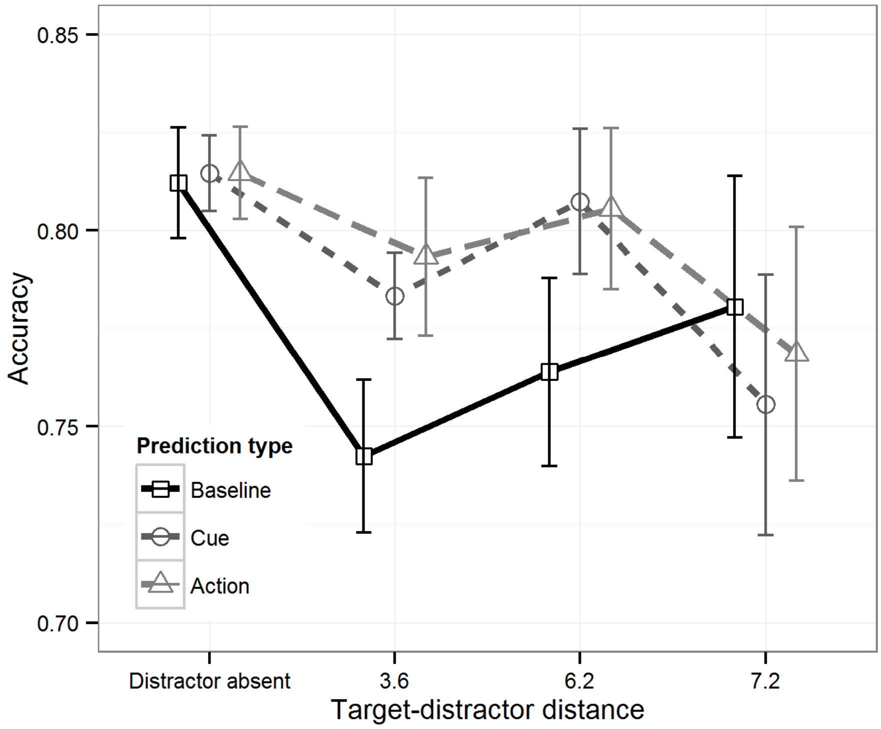

***Figure 4-4. Accuracy per prediction type condition and target-distractor distance.** Error bars show 95% confidence intervals corrected for within-subject designs.*

4.7.2 Contrasts between prediction processes

***Table 4-4. Contrasts between the baseline, action and cue prediction conditions to isolate action-effect prediction processing.** According to Hughes et al. 2013.*

Condition or contrast type	Temporal prediction	Temporal control	Non-motor identity prediction	Motor identity prediction
Baseline	√	√		
Cue	√	√	√	
Action	√	√	√	√
Contrast C-B			√	
Contrast A-C				√

5 General Discussion

The sense of agency stands for a variety of diverse phenomena connected with us being agents, having a body and performing goal-directed actions with consequences in the world and being aware of all of that. The term can refer to various abilities, such as being able to say whether some event was the result of my action, performed by my body, what it was and why I did it. These reports can be grouped under the term of judgment of agency. But it can also refer to various phenomenal experiences, such as what it is like to own a body, to perform an action with some intention, accidentally, or under compulsion, and to control the action fluently or in the face of external disturbances, grouped under the term of feeling or experience of agency. Or at least that is a common opinion in the field. While whether we possess the abilities, which of them and how they work is an objective fact, in principle discoverable by science, whether we really commonly or at least sometimes experience such phenomenal experiences, which of them, in what specific form and in which situation is much harder to tell. I personally do not feel any special experience of voluntary action or of owning my fingers when I am voluntarily typing these words or at least I do not think so. I am merely focusing on the writing and thinking about how to best express my ideas. I can pause and in retrospect reflect that what I am doing is voluntary, goal-oriented, and so on, that I take myself to be a free agent. Maybe there is no special experience common to voluntary action, but maybe it is just too subtle for me to be able to recognize it. It has been argued that people are in general not good at introspecting own experiences (Schwitzgebel, 2008). Maybe training in phenomenological reflection could help me (Gallagher & Zahavi, 2008). Or maybe there are just the experiences of something going wrong, such as when I want to raise my hand in the morning but am unable to do so because it is paralyzed from me sleeping on it. Or when I am used to cause an automatic escalator to start moving by stepping on it, but this time my action does not produce this effect. In both cases I experience a very peculiar feeling, but for me it is just a feeling of strangeness, not of something that I could individuate as a feeling of lack of control or lack of agency in the way I can individuate e.g. feeling of hunger. Similarly, in the various pathological cases (section 1.2), there may be a special feeling of lack of agency or control or ownership, but this does not need to imply that there is a positive feeling of agency in normal cases. That is not to say that there are no conscious experiences related to agency, such as conscious experience of deliberation between two courses of action or of trying to figure out who caused some event. Similar argument is presented by Thor Grünbaum (2015). However, it seems common in the experimental literature to assume that people have various special experiences of agency which are salient enough that people can recognize changes in these experiences under the experimental manipulations and express the experience as a number. Or that we can infer the type or intensity of the experience from objective measure regardless of whether the participants are able to report the experience. That is the reason why I strongly believe that we need to occupy ourselves not only with more work on conceptual issues – what it is we want to investigate – but also on methodological issues – how we can and should investigate such matters. What methods we use determines what we end up investigating and what conclusions we can draw from our data.

The present thesis aims to contribute to these methodological issues. After sketching what the general challenge is (section 1.3) and introducing the basic concepts and theories (section 1.4) I analyze in some detail selected prominent studies that use explicit and implicit methods and identify some general issues connected with them (chapter 2). Suggestions of methodological recommendations for future research will follow (section 5.3). This thesis also contains two empirical studies, one related to explicit methods and one to implicit. In the former study (chapter 3), we investigate the reliability of reports on determinants of own behavior and of reports on experience of control. In the latter study (chapter 4), we investigate whether we perceive and attend to predictable sensory consequences of our own actions differently than to predictable external events, as is assumed by the motor control-based theories of the sense of agency (section 1.4.1) and sensorimotor implicit measures (section 2.2.2) or whether the effect can be better explained by more general mechanisms (such as in section 1.4.4). The conclusions of the empirical studies will be discussed first.

5.1 Are all reports on agency reliable?

In the study *Metacognition of determinants of behavior: Learning to know more that we can tell* (chapter 3), we investigate the metacognitive and introspective abilities of people regarding their actions in a variation on the classical stop-signal task. On some trials people were asked to decide between responding and non-responding and then to report whether their action (or a lack of action) was the result of their deliberation or whether it was an automatic, reflexive response. Because the instructions and the concept of decision in particular can be ambiguous or difficult to understand, we took care to explain what is wanted from the participants in detail and provide sufficient practice. Participants indicated good understanding of the task before the experiment and also after the end of the experiment. We further excluded participants whose performance could signify lack of understanding of the instructions. Nevertheless, we found that many participants committed a substantial amount of false alarms, i.e., claiming to have made a deliberated response, although their reaction times strongly suggested that the response was in fact automatic. Next, we provided half of the participants with a metacognitive training, during which they received feedback messages after their decision judgments based on their response times, while the other half of participants received random feedback. Using fuzzy signal detection theory we found that the training increased the metacognitive sensitivity of the trained participants but did not influence their response bias or other aspects of performance. Metacognitive reports on the determinants of one’s behavior were often unreliable, suggestive of post-hoc rationalization (section 1.4.2), presumably because we do not have good access to such processes, especially in the artificial situation of our experiment, in which participants were under time pressure and had no personal reasons for making their decisions. Still, participants were able to improve their metacognition to some degree, suggesting that the reliability of explicit reports on agency can be improved. However, our case was special in that we were able to use the reaction times as a sort of implicit measure to make inferences about participants’ cognitive processes and provide feedback based on that. Although the metacognitive judgments were often unreliable, participants tended to be well aware of the reliability, demonstrating rather good meta-metacognition. Participants also had insight into their performance in the primary task (letter discrimination) of the study, but did not have insight into their waiting (response-postponing) strategies, presumably because people were receiving feedback after each incorrect response on the primary task. Moreover, we found that ratings of feeling of control over actions were actually strongly related to the inverse of ratings of difficulty across conditions, regardless of the actual control that participants possessed, which we interpret as attribute substitution: internally translating an unusual question as being about something more readily available for report. Overall, people can provide reports of varying reliability depending on what the report is about, on how much practice (in the experiment and everyday life) and external validation about the reports one has received, and on the specific person, as there were large individual differences.

In chapter 2 I identified several methodological issues in the research on the sense of agency. I believe the presented study was designed not to suffer from them. While many studies ask for judgments on ambiguous, unusual concepts, we took great care to make sure that what we mean by decision was understood. I also came to the conclusions that there must be some clear fact of the matter to be answered in an experiment, otherwise one risks obtaining invalid answers. I would argue that in the presented study there is an objective fact of whether someone performed an automatic response or went through the processing stages required for an answer to count as a decision. Factuality of difference in processing can be evidenced by the bimodality of reaction times. From a subjective perspective of participants, they reported that it was clear to them about what their decision judgments were supposed to be. Moreover, that we were able to provide some kind of feedback about the correctness of the judgments and many participants indicated that the feedbacks were generally accurate also testifies about there having been some fact behind the judgments. This study moreover shows how such methodological issues are important, because when we asked about the feeling of control, the answers were probably in fact not about control over responses but about the perceived difficulty.

5.2 Can attention be an implicit cue for action-effect self-attribution?

In the study *Expect to be distracted: Prediction of salient distractor by action and cue attenuates its interference* (chapter 4) we investigated the influence of action-effect prediction on attentional processing of salient but task irrelevant distracting stimulus. The comparator model states that when performing an action a forward model uses an efference copy of our motor command and predicts the expected sensory consequences of this action and subtracts this prediction from the actual (estimate of) sensory consequences (section 1.4.1). The sensory effect of our own action is thus processed and perceived differently than sensory effect produced by somebody else, regardless of how well we can predict this externally produced effect, because we lack the efference information in that case. A common view (section 1.4.3) is that this difference in processing of action effects serves as a low-level cue for the sense of agency (specifically, self-attribution of sensory events). Similarly to the phenomenon of sensory attenuation, we hypothesized that prediction of an irrelevant distractor by action could reduce its interference, by attenuating the low-level sensory processing of or attentional orienting to the distractor. The intuition is that you do not need to pay attention to predictable, irrelevant, and potentially distracting side effects of your actions. As such, attenuation of the distractor could serve as an implicit measure of the engagement of the forward model and the comparator mechanism allegedly involved in action-effect self-attribution. However, apart from the action-effect prediction process (such as the forward model) there may be other processes responsible for the attenuation. That is why we designed our experiment according to the methodological framework by Hughes and colleagues (2013). Such a design allowed us to determine that the influence of the distractor was attenuated by mere knowledge of the distractor via a cue, with no (or non-detected) additional contribution of any action-effect prediction mechanism. We propose that the results can be explained by a more general predictive mechanism, such as proposed by the framework of predictive processing (section 1.4.4), which predicts that both types of predictive information (cue and action) in our study should produce the same attenuation, as they both predict the outcome with almost equal accuracy and precision.

With respect to the methodological issues discussed in chapter 2, this study also took the problems seriously and further showed how important proper methodology is. While from the obtained effect in action-prediction condition one might have concluded that there was e.g. a forward model at play, a comparison with the cue-prediction condition which carefully controlled for the presence of action was needed to make such a conclusion. However, this comparison showed that it likely was not a specific action-effect prediction mechanism that was responsible for the distractor-attenuation effect. Our study thus supports the conclusion of Hughes and colleagues that any sensorimotor implicit measure needs to be carefully validated to determine what processes it actually indexes. Moreover, we do not and cannot make any claims that participants felt higher experience of agency for the results of their actions in the action-prediction condition, nor that future findings of attenuated influence of distractor by an action will signify the presence or increase of feeling of agency. Asking the participants would in principle be the better way to learn about their subjective experiences, but we did not ask about any experience of agency or control, as the participants would likely not have a good understanding of the construct behind such question and, more importantly, neither would we, experimenters.

5.3 How can we investigate the sense of agency?

The chapter 2 was of a theoretical, methodological nature, reviewing how the sense of agency is commonly investigated and trying to illustrate how it could be investigated to arrive at more secure conclusions. The presented treatment of the implicit and explicit measures is not meant to give an exhaustive overview of all the possible methods that have been used to measure all the various phenomena grouped under the term “sense of agency” and related terms. Rather, I have tried to analyze selected exemplary studies to show that there are under-appreciated challenges in our endeavor to understand the sense of agency and that we need to be aware of them if we hope to progress.

First of all, there are in general acknowledged but in practice often neglected conceptual issues. The term “sense of agency” (and related terms) covers many phenomena, likely lacking any common “essence”. Gallagher (2012) arrives in his analysis to a conclusion that not only is there an ambiguity in the concept of the sense of agency (definitions), but in the sense of agency itself (phenomenology), giving the impression that the phenomenon is indeterminate, as if what I experience depends on your point of view, and that we cannot formulate a coherent theory (conceptualization) of it. However, using the same term (or several similar terms) across many different studies can lead to implicit essentialism, the appearance that there is one thing about which all the studies inform us, while they actually study very diverse and possibly often completely distinct phenomena. That is why I suggest using less general (because it is unclear whether the term “sense of agency” really is general, i.e., superordinate to all the specific phenomena), more specific terms, describing the investigated phenomenon in behavioral terms if possible. For instance, instead of “sense of agency” use “self-attribution of sensory event” if the experiment requires participants to say who produced a certain effect. Instead of “experience of control” use “judgment of physical effort”, if we are dealing with a paradigm manipulating physical difficulty. Instead of “sense of ownership” use “ability to visually identify one’s own body parts”, if a feedback distortion paradigm is utilized. Instead of “representation of agency” use e.g. “representation of action plan” to elucidate what specific role the construct plays in your theory. Such an approach should facilitate theory-building from empirical results and prevent confusions and potential obfuscation of meaning by jargon.[18]

We can see that conceptual issues are connected to methodological issues, that is, what you actually study depends on how you study it, and vice versa, what you want to study necessitates the usage of an appropriate method. It has repeatedly emerged in my analysis that there is a common conflation between the study of abilities and phenomenal experiences of people. This distinction is not clear-cut and there are complicated philosophical issues connected to it. However there are many occasions at which the distinction can be methodologically useful. Studies claiming to investigate experiences often in fact study various abilities and references to experiences are either superfluous or completely unwarranted. Sometimes the abilities are studied explicitly, such as an ability to recognize one’s own hand motion, but sometimes the studied ability is hidden behind a report aimed to answer a question about an experience and we thus can only guess what abilities were

[18] For example, “banning” a word and seeing whether it can be replaced by other descriptions can serve as a test for the presence of clear substance behind that specific usage of the word. See a discussion of the issue for example here:
http://lesswrong.com/lw/nu/taboo_your_words/
http://lesswrong.com/lw/nv/replace_the_symbol_with_the_substance/

employed in producing that report. I have used the concept of attribute substitution for this potential problem behind many explicit measures. When asking about unusual things, that are not part of daily communication, of ordinary language, about things that the participant cannot know or when there is no correct answer but one is still demanded, the likelihood of obtaining invalid (about something else than is asked) report increases. Asking about a certain experience, e.g., “sense of control”, does not guarantee that the answer will be about what the experimenter has in mind when theorizing about the “sense of control”. As we have seen, the report can be about the ability to perform a task well, to detect external disturbances, and many other things. How the term “control” can be misleading when assumed to mean the same thing in different studies and why it can be useful to replace it by behavioral descriptions can be illustrated on folk[19] intuitions about feeling in control over automated actions, such as driving a car. In one intuitive understanding, the more automated an action is, the more “mindless” it becomes, the less control over the behavior one experiences, the actions just happen. In another intuitive understanding, the more automated an action is, the better we can perform the action and automatically compensate for a variety of problems, the more controlled our behavior is.

Explicit reports can inform us about both abilities and experiences, depending on how we use them. If you are interested in an ability, design a task directly investigating the ability. In such tasks, there must be some fact about what a correct response is, such that you can give a feedback about the correctness of responses, and what the correct response is should vary from trial to trial. Otherwise you risk rationalizations, attribute substitutions, etcetera, on the side of the participants. If you are interested in experiences, design a situation in which some experiences are presumably experienced and ask about the experiences in a very careful way, such as in the EASE interview, if possible using multiple ways to ask about the construct you have in mind, like in the SOARS scale. When asking for trial-by-trial ratings of some experience, explain everything in detail, allow your participant to ask you for clarifications, discuss the experiment with them after it is over, and be cautious in interpreting the results, because in principle you cannot be sure what the ratings reflect. The experience should be salient enough to be accurately identifiable and reportable by the participants. It may not be always meaningful to rate an experience on a continuous scale, as if it is a quantifiable, unidimensional variable, although sometimes this can be so. Instead, we can use a questionnaire and factor analysis in a pilot study to identify whether there may be more dimensions to the experience. We can also ask participants to come up with their own rating categories, such as

[19] or at least the author’s, as an example of possible intuitions of participants in such studies

the ordinal categories (No experience, Brief glimpse, Almost clear experience, and Clear experience) of the perceptual awareness scale (Ramsøy & Overgaard, 2004).

Implicit methods, in turn, can inform us mostly about the neuro-cognitive processes underlying certain abilities. This is a common practice in cognitive sciences and we have employed it in the study reported in chapter 3 to infer the presence or absence of a group of processes. In general, depending on the strength of our knowledge of the likelihood p(M|P) that an implicit measure M is produced by a given process P and on our knowledge about the base rate p(M) of obtaining the measure regardless of the engagement of the process, we can increase or decrease the strength of our prior belief p(P) that the process is engaged in the situation in which we measure M, arriving at a posterior probability p(P|M) (section 2.2.1). Such a reverse inference is necessarily uncertain and does not seem to be easy to perform for the current sensorimotor implicit measures (section 2.2.2). But in principle – although we have seen in chapter 4 that in practice it is rather challenging – we can reach the stage at which we can infer the presence or absence of some generally specified process, e.g. of causality detection. Going to a finer level of granularity, i.e., inferring a specific state of the process, then requires not only a wealth of experimental data – more research on the fundamental mechanisms of perception, prediction, attention, action, decision-making and so on – but also an ontology (taxonomy, conceptualization) of cognitive processes.

Although inferring a process from an objective measure is in principle possible, inferring a subjective experience presents a whole new level of added complexity. Indeed, such a project has long been the goal of the science of consciousness, with intense discussions across philosophy and empirical sciences on how such an inference could be possible given the inherent privacy of the subjective experience. It is not my intention to claim that such a project cannot succeed, but merely to highlight the far-from-trivial difficulties connected with it at the moment. I am currently skeptical about the possibility to infer experiences of agency from implicit measures like the intentional (or temporal) binding with much certainty and resolution. Even if we can reliably infer the engagement of e.g. causality processing, we cannot a priori assume that the process needs to be connected with some phenomenology. But even if we grant that there is such phenomenology, when we record a change in the measure, all that we can safely conclude is that there was a change in the process as well, but not that there was a change in the experience, because while the relation of supervenience of experience on physical processes is easy to accept if we are materialists, the relation of identity (one-to-one, *any* change in experience-constituting process entails a change in the experience) is much stronger and cannot be simply assumed by the experimenters, without even stating this assumption and providing good arguments for it (for more see section 2.2.1). We should therefore refrain from making claims about subjective experiences (see examples in section 2.2.2) based on at least the current sensorimotor implicit measures for the sense of agency.

Instead of assuming such a strong connection between processes and experiences, we can establish an objective measure for some experiencing by correlating the measure with explicit reports, where the measure can be as complex as a snapshot of brain activity, like in the search for the neural correlates of consciousness. This project faces many obstacles and a proper treatment of it is far beyond the scope of the present thesis. Some of the obstacles concern the problem of reverse inference and a need for phenomenal ontology and subjects trained in individuating experiences and providing reports according to the ontology. Perhaps one day we may be able – thanks to the well-trained subjects and advances in neuroimaging technology – to decode experiences from other people who are not able or willing to identify and report the experiences. But the problem of uncertainty of the inference will likely remain.

It still seems as the best option to learn what someone is experiencing to ask her or him. We have seen in chapter 3 that not all reports and not from all participants are always reliable. But we have also seen that which reports and under what conditions are reliable can be often investigated empirically and that there are better and worse ways of inquiring about experiences. The quality of reports likely depends on how salient the experience is and on how commonly the experience figures in ordinary conversations. In contrast to that, the normal (non-pathological) feeling of agency may be too subtle to be reportable and the concept plays little if any role in our ordinary language. It may very well be a theorist's invention with no need of explanation and also not explaining any known facts (Grünbaum, 2015).

However, what seems to demand an explanation and should be taken seriously are the peculiar reported experiences and deficits in abilities in pathological conditions such as delusions of control, which have motivated the research on the sense of agency in the first place. In these disorders the experiences are presumably salient enough for reports of informative value. Whether a patient feels a hand under his or her control or as having a mind of its own is a very different matter from whether a healthy participant "feels in control" over a color patch on a screen to a degree rated as 48 or 50 or whether the participant has a temporal bias of 150 or 200 ms when judging onsets of actions and effects. Even though one has to be even more careful about interpreting reports from patients than from healthy participants, more clinical research on experiencing in patients can be the way to substantially advance our understanding of the experiences and the cognitive and neural processes underlying them, and ultimately, helping the patients.

In conclusion, investigating the sense of agency by explicit and implicit methods is challenging, as is the whole science of consciousness. Plurality of approaches and interdisciplinary work are required, ranging from philosophy, psychological sciences, neurosciences, to computational modelling. However, we do not need only more empirical results and theoretical and conceptual progress, but also methodological progress, to know how to answer which questions and what conclusions we can safely draw from our results. By properly employing various explicit and implicit methods, we can advance our understanding of the experiences, abilities, and neuro-cognitive processes connected with the notion of sense of agency. I hope that the present work can be beneficial in this respect to researchers studying not only the sense of agency but consciousness in general.

6 References

Adams, R. A., Stephan, K. E., Brown, H. R., Frith, C. D., & Friston, K. J. (2013). The computational anatomy of psychosis. *Frontiers in Psychiatry*, *4*, 47. http://doi.org/10.3389/fpsyt.2013.00047

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(8), 2960–6. http://doi.org/10.1523/JNEUROSCI.3730-10.2010

Apps, M. A. J., & Tsakiris, M. (2014). The free-energy self: A predictive coding account of self-recognition. *Neuroscience and Biobehavioral Reviews*, *41*(0), 85–97. http://doi.org/10.1016/j.neubiorev.2013.01.029

Arita, J. T., Carlisle, N. B., & Woodman, G. F. (2012). Templates for rejection: configuring attention to ignore task-irrelevant features. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(3), 580–4. http://doi.org/10.1037/a0027885

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception & Psychophysics*, *55*(5), 485–96.

Baess, P., Horváth, J., Jacobsen, T., & Schröger, E. (2011). Selective suppression of self-initiated sounds in an auditory stream: An ERP study. *Psychophysiology*, *48*(9), 1276–83. http://doi.org/10.1111/j.1469-8986.2011.01196.x

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215.

Barlas, Z., & Obhi, S. S. (2013). Freedom, choice, and the sense of agency. *Frontiers in Human Neuroscience*, *7*(August), 514. http://doi.org/10.3389/fnhum.2013.00514

Bear, A., & Bloom, P. (2016). A Simple Task Uncovers a Postdictive Illusion of Choice. *Psychological Science*, *0*(0), 0. http://doi.org/10.1177/0956797616641943

Beck, V., Luck, S., & Hollingworth, A. (2011). The Implementation of an Exclusionary Attentional Template: Direct Versus Indirect Cueing. *Journal of Vision*, *11*(11), 1309–1309. http://doi.org/10.1167/11.11.1309

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, *74*(3), 183–200.

Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (1999). Spatio-temporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience*, *11*(5), 551–559.

Blakemore, S. J., Wolpert, D. M., & Frith, C. D. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, *1*(7), 635–640. http://doi.org/10.1038/2870

Blaney, P. H. (2009). Paranoid and delusional disorders. In *Oxford textbook of psychopathology (2nd ed.).* (pp. 361–396).

Block, N. (1995). On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, *18*, 227–287.

Boucsein, W. (2012). *Electrodermal Activity. Springer Science & Business Media.* (Second Edi).

Boston, MA: Springer US. http://doi.org/10.1007/978-1-4614-1126-0

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414. http://doi.org/10.1037/a0026450

Brischoux, F., & Angelier, F. (2015). Academia's never-ending selection for productivity. *Scientometrics*, *103*(1), 333–336. http://doi.org/10.1007/s11192-015-1534-5

Brown, H., Adams, R. a, Parees, I., Edwards, M., & Friston, K. (2013). Active inference, sensory attenuation and illusions. *Cognitive Processing*, *14*(4), 411–27. http://doi.org/10.1007/s10339-013-0571-3

Brown, H., Friston, K., & Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in Psychology*, *2*(September), 218. http://doi.org/10.3389/fpsyg.2011.00218

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822–848. http://doi.org/10.1037/0022-3514.84.4.822

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*, 25. http://doi.org/10.3389/fnhum.2010.00025

Buckolz, E., Guy, S., Khan, M., & Lawrence, G. (2006). Can the location negative priming process operate in a proactive manner? *Psychological Research*, *70*(3), 218–27. http://doi.org/10.1007/s00426-004-0202-9

Buehner, M. J. (2012). Understanding the Past, Predicting the Future: Causation, Not Intentional Action, Is the Root of Temporal Binding. *Psychological Science*, *23*, 1490–1497. http://doi.org/10.1177/0956797612444612

Buehner, M. J., & Humphreys, G. R. (2009). Causal Binding of Actions to Their Effects. *Psychological Science*, *20*(10), 1221–1228. http://doi.org/10.1111/j.1467-9280.2009.02435.x

Cardoso-Leite, P., Mamassian, P., Schütz-Bosbach, S., & Waszak, F. (2010). A new look at sensory attenuation. Action-effect anticipation affects sensitivity, not response bias. *Psychological Science*, *21*(12), 1740–5. http://doi.org/10.1177/0956797610389187

Carruthers, G. (2012). The case for the comparator model as an explanation of the sense of agency and its breakdowns. *Consciousness and Cognition*, *21*(1), 30-45–8. http://doi.org/10.1016/j.concog.2010.08.005

Caspar, E. A., Christensen, J. F., Cleeremans, A., & Haggard, P. (2016). Coercion Changes the Sense of Agency in the Human Brain. *Current Biology*, *26*(5), 585–592. http://doi.org/10.1016/j.cub.2015.12.067

Chadwick, P. K. (1993). The stepladder to the impossible: A first hand phenomenological account of a schizoaffective psychotic crisis. *Journal of Mental Health*, *2*(3), 239–250. http://doi.org/10.3109/09638239309003769

Chalmers, D. (1995). Facing up to the Problems of Consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Chambon, V., Moore, J. W., & Haggard, P. (2015). TMS stimulation over the inferior parietal cortex disrupts prospective sense of agency. *Brain Structure and Function*, *220*(6), 3627–3639. http://doi.org/10.1007/s00429-014-0878-6

Chambon, V., Wenke, D., Fleming, S. M., Prinz, W., & Haggard, P. (2013). An online neural substrate for sense of agency. *Cerebral Cortex*, *23*(5), 1031–1037.

http://doi.org/10.1093/cercor/bhs059

Chaminade, T., & Decety, J. (2002). Leader or follower? Involvement of the inferior parietal lobule in agency. *Neuroreport*, *13*(15), 1975–8.

Chao, H.-F. (2010). Top-down attentional control for distractor locations: the benefit of precuing distractor locations on target localization and discrimination. *Journal of Experimental Psychology. Human Perception and Performance*, *36*(2), 303–316. http://doi.org/10.1037/a0015790

Chen, C.-Y., Muggleton, N. G., Tzeng, O. J. L., Hung, D. L., & Juan, C.-H. (2009). Control of prepotent responses by the superior medial frontal cortex. *NeuroImage*, *44*(2), 537–45. http://doi.org/10.1016/j.neuroimage.2008.09.005

Chisholm, J. D., & Kingstone, A. (2014). Knowing and avoiding: the influence of distractor awareness on oculomotor capture. *Attention, Perception & Psychophysics*, *76*(5), 1258–64. http://doi.org/10.3758/s13414-014-0662-y

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. http://doi.org/10.1017/S0140525X12000477

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. *Statistical Power Analysis for the Behavioral Sciences* (2nd Revise, Vol. 2nd). Hillsdale, NJ: Lawrence Erlbaum Associates Inc. http://doi.org/10.1234/12345678

Cohen, M. A., & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15*(8), 358–64. http://doi.org/10.1016/j.tics.2011.06.008

Daprati, E., Franck, N., Georgieff, N., Proust, J., Pacherie, E., Dalery, J., & Jeannerod, M. (1997). Looking for the agent: an investigation into consciousness of action and self-consciousness in schizophrenic patients. *Cognition*, *65*(1), 71–86. http://doi.org/10.1016/S0010-0277(97)00039-5

David, N., Newen, A., & Vogeley, K. (2008). The “sense of agency” and its underlying cognitive and neural mechanisms. *Consciousness and Cognition*, *17*(2), 523–34. http://doi.org/10.1016/j.concog.2008.03.004

De Jong, R., Coles, M. G., & Logan, G. D. (1995). Strategies and mechanisms in nonselective and selective inhibitory motor control. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(3), 498–511.

De Jong, R., Coles, M. G., Logan, G. D., & Gratton, G. (1990). In search of the point of no return: the control of response processes. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(1), 164–182.

Decaix, C., Siéroff, E., & Bartolomeo, P. (2002). How Voluntary is “Voluntary” Orienting of Attention? *Cortex*, *38*(5), 841–845. http://doi.org/10.1016/S0010-9452(08)70053-4

den Ouden, H. E. M., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*(December), 548. http://doi.org/10.3389/fpsyg.2012.00548

Dennett, D. C. (2003). Who’s on first? Heterophenomenology explained. *Journal of Consciousness Studies*, *10*(9–10), 19–30.

Dennett, D. C. (2007). Heterophenomenology reconsidered. *Phenomenology and the Cognitive*

Sciences, *6*(1), 247–270.

Desantis, A., Hughes, G., & Waszak, F. (2012). Intentional binding is driven by the mere presence of an action and not by motor prediction. *PloS One*, *7*(1), e29557. http://doi.org/10.1371/journal.pone.0029557

Dewey, J. A., & Knoblich, G. (2014). Do Implicit and Explicit Measures of the Sense of Agency Measure the Same Thing? *PLoS ONE*, *9*(10), e110118. http://doi.org/10.1371/journal.pone.0110118

Dhawan, S., Deubel, H., & Jonikaitis, D. (2013). Inhibition of saccades elicits attentional suppression. *Journal of Vision*, *13*(6), 1–12. http://doi.org/10.1167/13.6.9

Dogge, M., Schaap, M., Custers, R., Wegner, D. M., & Aarts, H. (2012). When moving without volition: Implied self-causation enhances binding strength between involuntary actions and effects. *Consciousness and Cognition*, *21*(1), 501–506. http://doi.org/10.1016/j.concog.2011.10.014

Duncan, J. (1985). Visual search and visual attention. In M. Posner & O. Marin (Eds.), *Attention and performance XI* (pp. 85–106). NJ: Erlbaum: Hillsdale.

Eimer, M., & Kiss, M. (2008). Involuntary attentional capture is determined by task set: evidence from event-related brain potentials. *Journal of Cognitive Neuroscience*, *20*(8), 1423–33. http://doi.org/10.1162/jocn.2008.20099

Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychological Review*, *87*(3), 215–251. http://doi.org/10.1037/0033-295X.87.3.215

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

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(January), 429–433. http://doi.org/10.1038/415429a

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: A positron emission tomography study. *NeuroImage*, *18*(2), 324–333. http://doi.org/10.1016/S1053-8119(02)00041-1

Farrer, C., Franck, N., Paillard, J., & Jeannerod, M. (2003). The role of proprioception in action recognition. *Consciousness and Cognition*, *12*(4), 609–619. http://doi.org/10.1016/S1053-8100(03)00047-3

Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs another person as being the cause of an action: the neural correlates of the experience of agency. *NeuroImage*, *15*(3), 596–603. http://doi.org/10.1006/nimg.2001.1009

Farrer, C., Valentin, G., & Hupé, J. M. (2013). The time windows of the sense of agency. *Consciousness and Cognition*, *22*(4), 1431–1441. http://doi.org/10.1016/j.concog.2013.09.010

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience*, *4*, 215. http://doi.org/10.3389/fnhum.2010.00215

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. http://doi.org/10.1098/rstb.2012.0021

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human*

Neuroscience, *8*, 443. http://doi.org/10.3389/fnhum.2014.00443

Fleming, S. M., Weil, R. S., Nagy, Z., Dolan, R. J., & Rees, G. (2010). Relating introspective accuracy to individual differences in brain structure. *Science*, *329*(5998), 1541–3. http://doi.org/10.1126/science.1191883

Fletcher, P. C., & Frith, C. D. (2009). Perceiving is believing: a Bayesian approach to explaining the positive symptoms of schizophrenia. *Nature Reviews. Neuroscience*, *10*(1), 48–58. http://doi.org/10.1038/nrn2536

Fotopoulou, A., Tsakiris, M., Haggard, P., Vagopoulou, A., Rudd, A., & Kopelman, M. (2008). The role of motor intention in motor awareness: An experimental study on anosognosia for hemiplegia. *Brain*, *131*(12), 3432–3442. http://doi.org/10.1093/brain/awn225

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *360*(1456), 815–36. http://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews. Neuroscience*, *11*(2), 127–38. http://doi.org/10.1038/nrn2787

Friston, K. (2011). What is optimal about motor control? *Neuron*, *72*(3), 488–498. http://doi.org/10.1016/j.neuron.2011.10.018

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230–1233. http://doi.org/10.1016/j.neuroimage.2011.10.004

Frith, C. D. (1992). *The cognitive neuropsychology of schizophrenia*. Hove, UK: Erlbaum.

Frith, C. D. (2005). The self in action: Lessons from delusions of control. *Consciousness and Cognition*, *14*(4), 752–770. http://doi.org/10.1016/j.concog.2005.04.002

Frith, C. D. (2012). Explaining delusions of control: the comparator model 20 years on. *Consciousness and Cognition*, *21*(1), 52–4. http://doi.org/10.1016/j.concog.2011.06.010

Frith, C. D., Blakemore, S.-J., & Wolpert, D. M. (2000). Explaining the symptoms of schizophrenia: abnormalities in the awareness of action. *Brain Research. Brain Research Reviews*, *31*(2–3), 357–63.

Frith, C. D., & Friston, K. J. (2013). False perceptions and false beliefs: understanding schizophrenia. In A. Battro, S. Dehaene, & W. Singer (Eds.), *Neuroscience and the Human Person: New Perspectives on Human Activities* (pp. 1–15). Vatican City: Pontifical Academy of Sciences.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*, *4*(1), 14–21.

Gallagher, S. (2007). The Natural Philosophy of Agency. *Philosophy Compass*, *2*(2), 347–357. http://doi.org/10.1111/j.1747-9991.2007.00067.x

Gallagher, S. (2012). Multiple aspects in the sense of agency. *New Ideas in Psychology*, *30*(1), 15–31. http://doi.org/10.1016/j.newideapsych.2010.03.003

Gallagher, S., & Zahavi, D. (2008). *The Phenomenological Mind*. New York: Routledge.

Gaspar, J. M., & McDonald, J. J. (2014). Suppression of Salient Objects Prevents Distraction in Visual Search. *Journal of Neuroscience*, *34*(16), 5658–5666. http://doi.org/10.1523/JNEUROSCI.4161-13.2014

Gentsch, A., & Schütz-Bosbach, S. (2011). I did it: unconscious expectation of sensory consequences modulates the experience of self-agency and its functional signature. *Journal of Cognitive Neuroscience*, *23*(12), 3817–28. http://doi.org/10.1162/jocn_a_00012

Gentsch, A., & Synofzik, M. (2014). Affective coding: the emotional dimension of agency. *Frontiers in Human Neuroscience*, *8*, 608. http://doi.org/10.3389/fnhum.2014.00608

Gibson, B. S., & Jiang, Y. (1998). Surprise! An Unexpected Color Singleton Does Not Capture Attention in Visual Search. *Psychological Science*, *9*(3), 176–182. http://doi.org/10.1111/1467-9280.00034

Graham, G., & Stephens, G. L. (1994). Mind and mine. In G. Graham & G. L. Stephens (Eds.), *Philosophical Psychopathology* (pp. 91–109). Cambridge, MA: The MIT Press.

Graves, P. (2010). *Consumer.ology: The Market Research Myth, the Truth about Consumers, and the Psychology of Shopping*. London: Nicholas Brealey Publishing.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grünbaum, T. (2015). The feeling of agency hypothesis: a critique. *Synthese*, *192*(10), 3313–3337. http://doi.org/10.1007/s11229-015-0704-6

Haering, C., & Kiesel, A. (2014). Intentional Binding is independent of the validity of the action effect's identity. *Acta Psychologica*, *152*, 109–119. http://doi.org/10.1016/j.actpsy.2014.07.015

Haggard, P. (2008). Human volition: towards a neuroscience of will. *Nature Reviews. Neuroscience*, *9*(12), 934–46. http://doi.org/10.1038/nrn2497

Haggard, P., & Chambon, V. (2012). Sense of agency. *Current Biology : CB*, *22*(10), R390-2. http://doi.org/10.1016/j.cub.2012.02.040

Haggard, P., Clark, S., & Kalogeras, J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience*, *5*(4), 382–5. http://doi.org/10.1038/nn827

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. http://doi.org/10.3758/BF03203619

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. http://doi.org/10.1038/nrn1931

Helmholtz, H. von. (1867). *Handbuch der Physiologischen Optik*. Leipzig: L. Voss.

Herwig, A., & Waszak, F. (2012). Action-effect bindings and ideomotor learning in intention- and stimulus-based actions. *Frontiers in Psychology*, *3*(October 2012), 444. http://doi.org/10.3389/fpsyg.2012.00444

Hickey, C., McDonald, J. J., & Theeuwes, J. (2006). Electrophysiological evidence of the capture of visual attention. *Journal of Cognitive Neuroscience*, *18*(4), 604–13. http://doi.org/10.1162/jocn.2006.18.4.604

Hillyard, S., Vogel, E., & Luck, S. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1257–127.

Hitlin, S., & Elder, G. H. J. (2007). Time , Self , and the Curiously Abstract Concept of Agency.

Sociological Theory, *25*(2), 170–191.

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*(April), 96. http://doi.org/10.3389/fpsyg.2012.00096

Hommel, B., Pratt, J., Colzato, L., & Godijn, R. (2001). Symbolic control of visual attention. *Psychological Science*, *12*(5), 360–365.

Hopf, J.-M., Boehler, C. N., Luck, S. J., Tsotsos, J. K., Heinze, H.-J., & Schoenfeld, M. A. (2006). Direct neurophysiological evidence for spatial suppression surrounding the focus of attention in vision. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(4), 1053–8. http://doi.org/10.1073/pnas.0507746103

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychological Bulletin*, *139*(1), 133–151. http://doi.org/10.1037/a0028566

Jeannerod, M. (2006). *Motor Cognition: What Actions Tell the Self.* Oxford: Oxford University Press.

Jiang, J., Summerfield, C., & Egner, T. (2013). Attention sharpens the distinction between expected and unexpected percepts in the visual brain. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *33*(47), 18438–47. http://doi.org/10.1523/JNEUROSCI.3308-13.2013

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science*, *310*(5745), 116–119. http://doi.org/10.1126/science.1111709

Kahneman, D. (2011). *Thinking, Fast and Slow*. *Book* (Vol. 1). New York: Farrar, Straus & Giroux Inc.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*. New York: Cambridge University Press. http://doi.org/10.1038/2251090a0

Kihlstrom, J. F. (2008). The automaticity juggernaut—or, are we automatons after all. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), *Are we free? Psychology and free will* (pp. 155–180). New York: Oxford University Press.

Kiss, M., Grubert, A., Petersen, A., & Eimer, M. (2012). Attentional capture by salient distractors during visual search is determined by temporal task demands. *Journal of Cognitive Neuroscience*, *24*(3), 749–59. http://doi.org/10.1162/jocn_a_00127

Knoblich, G., & Sebanz, N. (2005). Agency in the face of error. *Trends in Cognitive Sciences*, *9*(6), 259–61. http://doi.org/10.1016/j.tics.2005.04.006

Koch, A. I., Müller, H. J., & Zehetleitner, M. (2013). Distractors less salient than targets capture attention rather than producing non-spatial filtering costs. *Acta Psychologica*, *144*(1), 61–72. http://doi.org/10.1016/j.actpsy.2013.04.023

Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C., & de Lange, F. P. (2012). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex*, *22*(9), 2197–206. http://doi.org/10.1093/cercor/bhr310

Kornell, N. (2014). Where to draw the line on metacognition: A taxonomy of metacognitive cues.

Journal of Comparative Psychology, *128*(2), 160–2. http://doi.org/10.1037/a0036194

Kranick, S. M., Moore, J. W., Yusuf, N., Martinez, V. T., LaFaver, K., Edwards, M. J., … Voon, V. (2013). Action-effect binding is decreased in motor conversion disorder: Implications for sense of agency. *Movement Disorders*, *28*(8), 1110–1116. http://doi.org/10.1002/mds.25408

Kühn, S., & Brass, M. (2009). Retrospective construction of the judgement of free choice. *Consciousness and Cognition*, *18*(1), 12–21. http://doi.org/10.1016/j.concog.2008.09.007

Kühn, S., Nenchev, I., Haggard, P., Brass, M., Gallinat, J., & Voss, M. (2011). Whodunnit? Electrophysiological Correlates of Agency Judgements. *PLoS ONE*, *6*(12), e28657. http://doi.org/10.1371/journal.pone.0028657

Lahav, A., Makovski, T., & Tsal, Y. (2012). White bear everywhere: exploring the boundaries of the attentional white bear phenomenon. *Attention, Perception & Psychophysics*, *74*(4), 661–73. http://doi.org/10.3758/s13414-012-0275-2

Lamy, D., & Egeth, H. E. (2003). Attentional capture in singleton-detection and feature-search modes. *Journal of Experimental Psychology. Human Perception and Performance*, *29*(5), 1003–20. http://doi.org/10.1037/0096-1523.29.5.1003

Lange, K. (2011). The reduced N1 to self-generated tones: An effect of temporal predictability? *Psychophysiology*, *48*(8), 1088–1095. http://doi.org/10.1111/j.1469-8986.2010.01174.x

Leary, M. R. (2007). Motivational and emotional aspects of the self. *Annual Review of Psychology*, *58*, 317–344. http://doi.org/10.1146/annurev.psych.58.110405.085658

Leotti, L. A., Iyengar, S. S., & Ochsner, K. N. (2010). Born to choose: the origins and value of the need for control. *Trends in Cognitive Sciences*, *14*(10), 457–463. http://doi.org/10.1016/j.tics.2010.08.001

Levenson, H. (1981). Differentiating among internality, powerful others, and chance. In H. M. Lefcourt (Ed.), *Research with the Locus of Control* (Vol. 1, pp. 15–63). New York: Academic Press.

Lind, A., Hall, L., Breidegard, B., Balkenius, C., & Johansson, P. (2014). Speakers' Acceptance of Real-Time Speech Exchange Indicates That We Use Auditory Feedback to Specify the Meaning of What We Say. *Psychological Science*, *25*(6), 1198–1205. http://doi.org/10.1177/0956797614529797

Longo, M. R., Schüür, F., Kammers, M. P. M., Tsakiris, M., & Haggard, P. (2008). What is embodiment? A psychometric approach. *Cognition*, *107*(3), 978–98. http://doi.org/10.1016/j.cognition.2007.12.004

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–30. http://doi.org/10.1016/j.concog.2011.09.021

Marchetti, C., & Sala, S. Della. (1998). Disentangling the Alien and Anarchic Hand. *Cognitive Neuropsychiatry*, *3*(3), 191–207. http://doi.org/10.1080/135468098396143

Metcalfe, J., Eich, T. S., & Castel, A. D. (2010). Metacognition of agency across the lifespan. *Cognition*, *116*(2), 267–282. http://doi.org/10.1016/j.cognition.2010.05.009

Metcalfe, J., Eich, T. S., & Miele, D. B. (2013). Metacognition of agency: proximal action and distal outcome. *Experimental Brain Research*, *229*(3), 485–496. http://doi.org/10.1007/s00221-012-3371-6

Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology. General*, *136*(2), 184–99. http://doi.org/10.1037/0096-3445.136.2.184

Metcalfe, J., Van Snellenberg, J. X., DeRosse, P., Balsam, P., & Malhotra, A. K. (2012). Judgements of agency in schizophrenia: an impairment in autonoetic metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1391–1400. http://doi.org/10.1098/rstb.2012.0006

Miall, R. C., & Wolpert, D. M. (1996). Forward Models for Physiological Motor Control. *Neural Networks*.

Miele, D. B., Wager, T. D., Mitchell, J. P., & Metcalfe, J. (2011). Dissociating neural correlates of action monitoring and metacognition of agency. *Journal of Cognitive Neuroscience*, *23*(11), 3620–36. http://doi.org/10.1162/jocn_a_00052

Milgram, S. (1963). Behavioral Study Of Obedience. *Journal of Abnormal and Social Psychology*, *67*, 371–378.

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

Moore, J. W., Middleton, D., Haggard, P., & Fletcher, P. C. (2012). Exploring implicit and explicit aspects of sense of agency. *Consciousness and Cognition*, *21*(4), 1748–1753. http://doi.org/10.1016/j.concog.2012.10.005

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: A review. *Consciousness and Cognition*, *21*(1), 546–561. http://doi.org/10.1016/j.concog.2011.12.002

Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and Cognition*, *18*(4), 1056–64. http://doi.org/10.1016/j.concog.2009.05.004

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. http://doi.org/10.3758/s13414-012-0291-2

Mounts, J. R. W. (2005). Attentional selection: A salience-based competition for representation. *Perception & Psychophysics*, *67*(7), 1190–8.

Müller, H. J., Geyer, T., Zehetleitner, M., & Krummenacher, J. (2009). Attentional capture by salient color singleton distractors is modulated by top-down dimensional set. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(1), 1–16. http://doi.org/10.1037/0096-1523.35.1.1

Müller, H. J., Reimann, B., & Krummenacher, J. (2003). Visual search for singleton feature targets across dimensions: Stimulus- and expectancy-driven effects in dimensional weighting. *Journal of Experimental Psychology. Human Perception and Performance*, *29*(5), 1021–35. http://doi.org/10.1037/0096-1523.29.5.1021

Munneke, J., Van der Stigchel, S., & Theeuwes, J. (2008). Cueing the location of a distractor: an inhibitory mechanism of spatial attention? *Acta Psychologica*, *129*(1), 101–7. http://doi.org/10.1016/j.actpsy.2008.05.004

Nadelhoffer, T., Shepard, J., Nahmias, E., Sripada, C., & Ross, L. T. (2014). The free will inventory: Measuring beliefs about agency and responsibility. *Consciousness and Cognition*, *25*, 27–41. http://doi.org/10.1016/j.concog.2014.01.006

Nahmias, E., Coates, D. J., & Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy*, *31*(1), 214–242. http://doi.org/10.1111/j.1475-4975.2007.00158.x

Nahmias, E., Morris, S., Nadelhoffer, T., & Turner, J. (2005). Surveying Freedom: Folk Intuitions about free will and moral responsibility. *Philosophical Psychology*, *18*(5), 561–584. http://doi.org/10.1080/09515080500264180

Nielsen, T. I. (1963). Volition: A new experimental approach. *Scandinavian Journal of Psychology*, *4*(1), 225–230. http://doi.org/10.1111/j.1467-9450.1963.tb01326.x

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. http://doi.org/10.1037/0033-295X.84.3.231

Olivers, C. N. L. (2009). What drives memory-driven attentional capture? The effects of memory type, display type, and search type. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(5), 1275–91. http://doi.org/10.1037/a0013896

Pacherie, E. (2007). The Sense of Control and the Sense of Agency. *Psyche*, *13*(1), 1–30.

Pacherie, E. (2008). The phenomenology of action: a conceptual framework. *Cognition*, *107*(1), 179–217. http://doi.org/10.1016/j.cognition.2007.09.003

Parasuraman, R., Masalonis, A. J., & Hancock, P. A. (2000). Fuzzy Signal Detection Theory : Basic Postulates and Formulas for Analyzing Human and Machine Performance. *Human Factors*, *42*(4), 636–659.

Parnas, J., Møller, P., Kircher, T., Thalbitzer, J., Jansson, L., Handest, P., & Zahavi, D. (2005). EASE: Examination of anomalous self-experience. *Psychopathology*, *38*(5), 236–258. http://doi.org/10.1159/000088441

Paulhus, D. L., & Carey, J. M. (2011). The FAD-Plus: measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment*, *93*(1), 96–104. http://doi.org/10.1080/00223891.2010.528483

Peterson, S. A., & Gibson, T. N. (2011). Implicit attentional orienting in a target detection task with central cues. *Consciousness and Cognition*, *20*(4), 1532–47. http://doi.org/10.1016/j.concog.2011.07.004

Picard, F., & Friston, K. (2014). Predictions, perception, and a sense of self. *Neurology*, 1–7. http://doi.org/10.1212/WNL.0000000000000798

Pickering, M. J., & Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. *Trends in Cognitive Sciences*, *18*(9), 451–6. http://doi.org/10.1016/j.tics.2014.05.006

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63. http://doi.org/10.1016/j.tics.2005.12.004

Polito, V., Barnier, A. J., & Woody, E. Z. (2013). Developing the Sense of Agency Rating Scale (SOARS): An empirical measure of agency disruption in hypnosis. *Consciousness and Cognition*, *22*(3), 684–696. http://doi.org/10.1016/j.concog.2013.04.003

Poonian, S. K., & Cunnington, R. (2013). Intentional binding in self-made and observed actions. *Experimental Brain Research*, *229*(3), 419–427. http://doi.org/10.1007/s00221-013-3505-5

Poonian, S. K., McFadyen, J., Ogden, J., & Cunnington, R. (2015). Implicit Agency in Observed Actions: Evidence for N1 Suppression of Tones Caused by Self-made and Observed Actions.

Journal of Cognitive Neuroscience, *27*(4), 752–764. http://doi.org/10.1162/jocn_a_00745

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. http://doi.org/10.1080/00335558008248231

Raine, A. (1991). The SPQ: A Scale for the Assessment of Schizotypal Personality Based on DSM-III-R Criteria. *Schizophrenia Bulletin*, *17*(4), 555–564.

Ramsøy, T. Z., & Overgaard, M. (2004). Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, *3*(1), 1–23. http://doi.org/10.1023/B:PHEN.0000041900.30172.e8

Rangelov, D., Müller, H. J., & Zehetleitner, M. (2013). Visual search for feature singletons : Multiple mechanisms produce sequence effects in visual search. *Journal of Vision*, *13*(2), 1–16. http://doi.org/10.1167/13.3.22.doi

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. http://doi.org/10.1038/4580

Richters, D. P., & Eskew, R. T. (2009). Quantifying the effect of natural and arbitrary sensorimotor contingencies on chromatic judgments. *Journal of Vision*, *9*(4), 27.1-11. http://doi.org/10.1167/9.4.27

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, *80*(1), 1–28. http://doi.org/10.1017/CBO9781107415324.004

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. http://doi.org/10.3758/PBR.16.2.225

Roussel, C., Hughes, G., & Waszak, F. (2013). A preactivation account of sensory attenuation. *Neuropsychologia*, *51*(5), 922–9. http://doi.org/10.1016/j.neuropsychologia.2013.02.005

Ruff, C. C., & Driver, J. (2006). Attentional preparation for a lateralized visual distractor: behavioral and fMRI evidence. *Journal of Cognitive Neuroscience*, *18*(4), 522–38. http://doi.org/10.1162/jocn.2006.18.4.522

Ryle, G. (2000). *The Concept of Mind*. Chicago: The University of Chicago Press.

Sato, A. (2009). Both motor prediction and conceptual congruency between preview and action-effect contribute to explicit judgment of agency. *Cognition*, *110*(1), 74–83. http://doi.org/10.1016/j.cognition.2008.10.011

Sato, A., & Yasuda, A. (2005). Illusion of sense of self-agency: discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*, *94*(3), 241–55. http://doi.org/10.1016/j.cognition.2004.04.003

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnson (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). Windsor, UK: NFER-NELSON.

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *Philosophical Review*, *117*(2), 245–273. http://doi.org/10.1215/00318108-2007-037

Scott, S. H. (2004). Optimal feedback control and the neural basis of volitional motor control. *Nature Reviews. Neuroscience*, *5*(7), 532–46. http://doi.org/10.1038/nrn1427

Shore, D. I., Spence, C., & Klein, R. M. (2001). Visual prior entry. *Psychological Science : A Journal of the American Psychological Society / APS*, *12*(3), 205–212. http://doi.org/10.1111/1467-9280.00337

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology. General*, *117*(1), 34–50. http://doi.org/10.1037/0096-3445.117.1.34

Song, C., Kanai, R., Fleming, S. M., Weil, R. S., Schwarzkopf, D. S., & Rees, G. (2011). Relating inter-individual differences in metacognitive performance on different perceptual tasks. *Consciousness and Cognition*, *20*(4), 1787–1792. http://doi.org/10.1016/j.concog.2010.12.011

Spence, C., & Parise, C. (2010). Prior-entry: A review. *Consciousness and Cognition*, *19*(1), 364–379. http://doi.org/10.1016/j.concog.2009.12.001

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, *43*(6), 482–489. http://doi.org/10.1037/h0055479

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*(12), 1391–408. http://doi.org/10.1016/j.visres.2008.03.009

Strahan, D. (2016). Mid-career teachers' perceptions of self-guided professional growth: strengthening a sense of agency through collaboration. *Teacher Development*, *4530*(July), 1–15. http://doi.org/10.1080/13664530.2016.1190782

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–9. http://doi.org/10.1016/j.tics.2009.06.003

Synofzik, M., & Vosgerau, G. (2012). Beyond the comparator model. *Consciousness and Cognition*, *21*(1), 1–3. http://doi.org/10.1016/j.concog.2012.01.007

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Consciousness and Cognition*, *17*(1), 219–39. http://doi.org/10.1016/j.concog.2007.03.010

Synofzik, M., Vosgerau, G., & Voss, M. (2013). The experience of agency: An interplay between prediction and postdiction. *Frontiers in Psychology*, *4*(MAR), 1–8. http://doi.org/10.3389/fpsyg.2013.00127

Theeuwes, J. (1992). Perceptual selectivity for color and form. *Perception & Psychophysics*, *51*(6), 599–606. http://doi.org/10.3758/BF03211656

Theeuwes, J. (2010). Top-down and bottom-up control of visual selection. *Acta Psychologica*, *135*, 77–99. http://doi.org/10.1016/j.actpsy.2010.02.006

Theeuwes, J., Atchley, P., & Kramer, A. F. (2000). On the time course of top-down and bottom-up control of visual attention. In S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII* (pp. 105–125). Cambridge MA: MIT Press.

Titchener, E. (1908). *Lectures on the elementary psychology of feeling and attention*. New York: Macmillan.

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(25), 9118–23. http://doi.org/10.1523/JNEUROSCI.1425-11.2011

Tombu, M., & Tsotsos, J. K. (2008). Attending to orientation results in an inhibitory surround in orientation space. *Perception & Psychophysics*, *70*(1), 30–35. http://doi.org/10.3758/PP.70.1.30

Torbet, G., Schulze, D., Fiedler, A., & Reuter, B. (2015). Assessment of self-disorders in a non-clinical population: Reliability and association with schizotypy. *Psychiatry Research*, *228*(3), 857–865. http://doi.org/10.1016/j.psychres.2015.05.011

Tsakiris, M., & Haggard, P. (2005). The rubber hand illusion revisited: visuotactile integration and self-attribution. *J Exp Psychol Hum Percept Perform*, *31*(1), 80–91. http://doi.org/10.1037/0096-1523.31.1.80

Tsal, Y., & Makovski, T. (2006). The attentional white bear phenomenon: the mandatory allocation of attention to expected distractor locations. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(2), 351–63. http://doi.org/10.1037/0096-1523.32.2.351

Van den Bos, E., & Jeannerod, M. (2002). Sense of body and sense of action both contribute to self-recognition. *Cognition*, *85*(2), 177–187. http://doi.org/10.1016/S0010-0277(02)00100-2

Van Doorn, G., Hohwy, J., & Symmons, M. (2014). Can you tickle yourself if you swap bodies with someone else? *Consciousness and Cognition*, *23*, 1–11. http://doi.org/10.1016/j.concog.2013.10.009

van Gelder, T. (1998). The roles of philosophy in cognitive science. *Philosophical Psychology*, *11*(2), 117–136.

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *Bmj*, *351*(dec14_13), h6467. http://doi.org/10.1136/bmj.h6467

von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip - Wechselwirkungen zwischen Zentralnervensystem und Peripherie. *Die Naturwissenschaften*, *37*, 464–476. http://doi.org/10.1007/BF00622503

Vroomen, J., & Stekelenburg, J. J. (2010). Visual Anticipatory Information Modulates Multisensory Interactions of Artificial Audiovisual Stimuli. *Journal of Cognitive Neuroscience*, *22*(7), 1583–1596. http://doi.org/10.1162/jocn.2009.21308

Walach, H., Buchheld, N., Buttenmüller, V., Kleinknecht, N., & Schmidt, S. (2006). Measuring mindfulness—the Freiburg Mindfulness Inventory (FMI). *Personality and Individual Differences*, *40*(8), 1543–1555. http://doi.org/10.1016/j.paid.2005.11.025

Waszak, F., Cardoso-Leite, P., & Hughes, G. (2012). Action effect anticipation: neurophysiological basis and functional consequences. *Neuroscience and Biobehavioral Reviews*, *36*(2), 943–59. http://doi.org/10.1016/j.neubiorev.2011.11.004

Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge: MIT Press.

Wegner, D. M. (2008). Self Is Magic. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), *Are We Free?: Psychology and Free Will* (pp. 226–247). New York: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780195189636.003.0011

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, *54*(7), 480–492. http://doi.org/10.1037/0003-066X.54.7.480

Weiss, C., Herwig, A., & Schütz-Bosbach, S. (2011). The self in action effects: selective attenuation of self-generated sounds. *Cognition*, *121*(2), 207–18.

http://doi.org/10.1016/j.cognition.2011.06.011

Wenke, D., Fleming, S. M., & Haggard, P. (2010). Subliminal priming of actions influences sense of control over effects of action. *Cognition*, *115*(1), 26–38. http://doi.org/10.1016/j.cognition.2009.10.016

Wolpert, D. M., & Flanagan, J. (2001). Motor prediction. *Current Biology*, *11*(18), R729–R732.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science*, *269*(5232), 1880–1882.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*(7–8), 1317–1329. http://doi.org/10.1016/S0893-6080(98)00066-5

Woodman, G. F., & Luck, S. J. (2007). Do the contents of visual working memory automatically influence attentional selection during visual search? *Journal of Experimental Psychology. Human Perception and Performance*, *33*(2), 363–77. http://doi.org/10.1037/0096-1523.33.2.363

Wykowska, A., & Schubö, A. (2010). On the temporal relation of top-down and bottom-up mechanisms during guidance of attention. *Journal of Cognitive Neuroscience*, *22*(4), 640–54. http://doi.org/10.1162/jocn.2009.21222

Wykowska, A., & Schubö, A. (2011). Irrelevant singletons in visual search do not capture attention but can produce nonspatial filtering costs. *Journal of Cognitive Neuroscience*, *23*(3), 645–60. http://doi.org/10.1162/jocn.2009.21390

Yantis, S. (1993). Stimulus-driven attentional capture. *Current Directions in Psychological Science*, *2*(5), 156–161.

Yoshie, M., & Haggard, P. (2013). Negative Emotional Outcomes Attenuate Sense of Agency over Voluntary Actions. *Current Biology*, *23*(20), 2028–2032. http://doi.org/10.1016/j.cub.2013.08.034

Zehetleitner, M., Koch, A. I., Goschy, H., & Müller, H. J. (2013). Salience-based selection: attentional capture by distractors less salient than the target. *PloS One*, *8*(1), e52595. http://doi.org/10.1371/journal.pone.0052595

7 Curriculum Vitae

Education

2012 – present Ludwig Maximilian University of Munich, Ph.D. program at the Graduate School of Systemic Neurosciences

2010 - 2012 University of Economics in Prague, Faculty of Informatics and Statistics, awarded Master's degree (Ing., with honors) in the program Applied Informatics, major in Cognitive Informatics, minor in Philosophy

09/2011 - 12/2011 Eötvös Loránd University, Budapest, Budapest Semester in Cognitive Science

2006 - 2010 University of Economics in Prague, Faculty of Informatics and Statistics, awarded Bachelor's degree (Bc.) in the program Applied Informatics, major in Informatics

01/2005 - 05/2005 Lawrence North High School, Indianapolis, IN, USA

1998 - 2006 Gymnázium Písek (grammar school)

Additional education

03/2016 *Munich Brain Course 2016 – Main topics: spinal cord, cerebellum and occipital lobe.* Hands-on intensive course in neuroanatomy, Munich, Germany

07/2015 Summer school *Computational Sensory-Motor Neuroscience (CoSMo 2015)*, Radboud University, Nijmegen, The Netherlands. Best group project award, presented at TCMC SfN 2015, Chicago, USA.

06/2015 Summer school *Consciousness and Decision Making*, Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Germany

03/2015 Workshop *Predictive brains*, on predictive coding, the free energy principle and neurophilosophy, VIU, Venice, Italy

06/2014 Spring school, International research group *Multisensory Perception for Action*, Wildbad Kreuth, Germany

04/2013 Neurophilosophy workshop *Argumentative theory of reasoning*, VIU, Venice, Italy

07/2012 Summer school *Problems of the Self*, Central European University in Budapest, Hungary

Professional experience

2015 National Institute of Mental Health, Prague, junior researcher

2012 – present Department of General and Experimental Psychology, LMU Munich, research project *Challenges of investigating the sense of agency with explicit and implicit methods*

2011 – 2012 Institute of Chemical Technology Prague, research project involving EEG and behavioral data analysis for a Master's thesis *Freedom of will and access to one's own intentions*

2010 - 2012 Clever Decision Ltd., Junior Business Intelligence consultant and developer

2009 Clever Decision Ltd., Bachelor's thesis *Business Intelligence Metadata Management in MS SQL Server 2005*

2008 Pados Ltd., Junior Java developer

Teaching experience

2014/2015 Practical tutorial *EEG and ERP methodology*, Master's program Neurocognitive Psychology, LMU Munich

2013/2014 Practical tutorial *Reaction-time and psychophysical methods*, Master's program Neurocognitive Psychology, LMU Munich

Publications

Kozáková, E., Havlíček, O., Bečev, O. (in press). Ukradené myšlenky a ovládané ruce (Stolen thoughts and externally controlled arms). National Institute of Mental Health, Prague.

Havlíček, O. (2013). Filosofický problém svobodné vůle ve světle vědeckých poznatků (Philosophical problem of free will in the light of scientific findings). *E-LOGOS*. WWW: <http://e-logos.vse.cz/index.php?article=343>

Selected academic presentations

Havlíček, O. (2015). *Predictive processing: A new paradigm for the cognitive sciences?* Talk at the National Institute of Mental Health, Prague.

Havlíček, O., Müller, H., Wykowska, A. (2014). *Expect to be distracted: Predicting salient distractors by action and cue*. (Talk and a poster at a symposium of the international research group "Multisensory perception for action", Wildbad Kreuth, Germany, and at a workshop "Predictive brains", VIU, Venice.)

Havlíček, O. (2013). *Naturalizace (problému) svobodné vůle.* (Naturalizing (the problem of) free will. Invited talk at the Palackého University in Olomouc, 2013 and at the University of West Bohemia in Pilsen, 2014)

Havlíček, O. (2013). *Compatibilism about freedom and arguing about reasons*. (Talk at a joint symposium of the Graduate School of Systemic Neurosciences and the Berlin School of Mind and Brain, VIU, Venice, 2013)

Havlíček, O. (2013). *Science and philosophy on agency and free will*. (Talk at a retreat of the Graduate School of Systemic Neurosciences, Chiemsee, 2013)

Havlíček, O., Wykowska, A., Sellmaier, S., Müller, H. (2013). *Sense of agency and action-perception links.* (Poster at an orientation week of the Graduate School of Systemic Neurosciences, Munich, 2013)

Havlíček, O. (2012). *Retrospective construction of the judgment of free choice.* (Poster at the summer school Problems of the Self, Budapest, 2012)

Research and IT skills

EEG (principles of EEG/ERP, extensive experience with data acquisition, analysis with BrainVision Analyzer and Matlab)

MRI (basic principles of MRI, fMRI, DTI, DCM, MVPA; hands-on experience with basic scanner operation, basic fMRI analysis with SPM, basic DTI analysis with FSL)

TMS (principles, basic hands-on experience, motor thresholds, co-registration with MRI scans)

Behavioral designs: E-Prime, OpenSesame (Python)

Statistical analysis: R, Matlab, SPSS

Programming and other languages: Java, C#, Matlab, VB, ASP.NET, SQL, XML

Language skills

Czech	Native speaker
English	Advanced (City & Guilds Level 2 Certificate in ESOL International (reading, writing and listening) Expert C1, 2005)
German	Intermediate (B1+)

8 List of Publications

Book chapter

Kozáková, E., Havlíček, O., Bečev, O. (in press). Ukradené myšlenky a ovládané ruce (Stolen thoughts and externally controlled arms). National Institute of Mental Health, Prague.

Peer-reviewed journal article

Havlíček, O. (2013). Filosofický problém svobodné vůle ve světle vědeckých poznatků (Philosophical problem of free will in the light of scientific findings). *E-LOGOS*. WWW: <http://e-logos.vse.cz/index.php?article=343>

9 Affidavit

Eidesstattliche Versicherung/Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation *The Challenges of Investigating the Sense of Agency by Explicit and Implicit Methods* selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation *The Challenges of Investigating the Sense of Agency by Explicit and Implicit Methods* is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, den Unterschrift
Munich, date signature

10 Declaration of Author Contributions

Chapter 1 (Introduction)

Ondřej Havlíček wrote the chapter.

Chapter 2 (Investigating the Sense of Agency)

Havlíček, O. (in preparation). The challenges of investigating the sense of agency by explicit and implicit methods.

OH performed the theoretical research and wrote the paper.

Chapter 3 (Study 1: Metacognition of Determinants of Behavior)

Havlíček, O., Brass, M., Cleeremans, A., Wykowska, A. (in preparation). Metacognition of determinants of behavior: Learning to know more that we can tell.

MB, OH, AW conceived and designed the study. OH performed the research and analyzed data. OH, AW, MB, AC wrote the paper.

Chapter 4 (Study 2: Expect to Be Distracted)

Havlíček, O., Müller, H.J., Wykowska, A. (in preparation). Expect to be distracted: Prediction of salient distractor by action and cue attenuates its interference.

OH, AW, HJM conceived and designed the study. OH performed the research and analyzed data. OH, AW, HJM wrote the paper.

Chapter 5 (General Discussion)

Ondřej Havlíček wrote the chapter.

Ondřej Havlíček (First author) Agnieszka Wykowska (Lab head)