# Empirical data in the philosophy of mind: free will, higher-order thought, and misrepresentation

Kirkeby-Hinrup, Asger

2017

*Document Version:*
Other version

# Empirical data in the philosophy of mind: free will, higher-order thought, and misrepresentation

Asger Kirkeby-Hinrup


LUND
UNIVERSITY

Coverphoto by Kalev Leetaru / Internet Archive Book Images



Faculty of Humanities
Department of Philosophy and Cognitive Science

ISBN 978 9188473400

Printed in Sweden by Media-Tryck, Lund University
Lund 2017




*For Cornelius, Emilie and Leonora*

# Table of contents

# Acknowledgements

Ingar Brinck, my supervisor, has without doubt had the biggest influence on my academic life of anybody. I am extremely thankful for your patience (I can imagine it has been tested) and relentless challenges to do better. Whenever I doubt my academic skills, I just read a few paragraphs of the first drafts I gave you to remind myself how far I have gotten by following your advice. In this respect, I am also indebted to Peter Gärdenfors and Erik Olsson who have served as my co-supervisors, and have provided valuable encouragement and academic feedback in that role. I also owe a special thanks to Peter for originally encouraging me to apply for the doctoral position.

The philosophy of mind reading group has been a coveted respite of mine and is one of the things I miss the most. I am extremely thankful to the members of the group – Andreas Falck, Fritz Gåvertsson, Patrizio Lo Presti and Oscar Ralsmark - for the many hours we have spent in conversation, particularly in the reading group sessions, staircases, and in front of Kungshuset. Somehow, the last two seemed just as conducive to detailed philosophical discussion as the former.

The members of the CogComLab research group, headed by Ingar, have provided valuable comments and insights on many topics. In addition to the aforementioned Lo Presti and Falck, these include Susanne Bernstrup, Åsa Harvard, Justine Jacot, Martin Jönsson and Elainie Madsen. Additionally, I want to thank the members of the NeuroInfoClub, hosted at the Department of Psychology by Emelie Stiernströmer. The NIC was an appreciated opportunity to learn from - and interact with - empirical researchers.

I am grateful to all the participants at the research seminars at theoretical philosophy over the years – too many to mention – not only for their comments and discussion on my presentations, but just as much for giving presentations of their own. I made an effort to attend every time, and listening to you all gave me a weekly reminder of how lucky I was to be working alongside so friendly and competent individuals. Actually, these characteristics apply to the all of the department at every the academic and administrative position, and my gratitude extends to all of you as well. In addition, I want to acknowledge my gratitude to Jan Hartman for giving me the opportunity to teach alongside him. Finally, I want to thank David Rosenthal for

taking the time to correspond at length with me. I will seek to pay forward this intellectual generosity to young academics throughout my career.

Thank you to Cathrine Felix and Paula Quinon for sharing friendship, worries and laughs. Casper Andersen and Arild Boes deserve a special mention for longtime friendship and support. The latter additionally for his eagerness always to listen to – and challenge – my philosophical predilections.

Getting this far would never have been possible without an inordinate amount of help and support from my family. Karen and Steen, Hans Jørgen and Johanne, Lars and Inge, and, of course, Ida. Thank you for everything.

Finally, my beloved wife: Thit. Thank you for willingly moving our family to another country to allow me to pursue philosophy. Thank you for supporting me. Thank you for putting up with me.

# List of papers

Paper I: Kirkeby-Hinrup, Asger. (2014a). Why the Rare Charles Bonnet Cases are not Evidence of Misrepresentation. *Journal of Philosophical Research*, 39, 301–308.

Paper II: Kirkeby-Hinrup, Asger. (2014b). How to Get Free Will from Positive Reinforcement. *SATS*, 15(1), 20–38.

Paper III: Kirkeby-Hinrup, Asger. (2015). How Choice Blindness Vindicates Wholeheartedness. *Organon F.*, 2(22), 199–210.

Paper IV: Kirkeby-Hinrup, Asger. (2016). Change Blindness and Misrepresentation. *Disputatio*, 8(42), 37–56.

Paper V: Brinck, Ingar & Kirkeby-Hinrup, Asger. (*In press*). Change blindness in higher-order thought: Misrepresentation or good enough? *Journal of Consciousness Studies*.

# Chapter 1 - Introduction

> Each, by historical accident, had inherited a particular way of looking at cognition and each had progressed far enough to recognize that the solution to some of its problems depended crucially on the solution of problems traditionally allocated to other disciplines.
>
> George A. Miller (2003, p. 143)

The theme of this dissertation is the interdisciplinary practices at the intersection between philosophy of mind and empirical sciences. Such interdisciplinary practice rests on the assumption that philosophy of mind intersects – with respect to the object of study – areas of the empirical sciences such as psychology, cognitive science, and neuroscience. The increase in interdisciplinary ventures involving philosophers of mind and empirical scientists over recent decades underscores the foresight of Miller, who in the quote above describes the birth of cognitive science in 1977. Originally, the field of cognitive science was conceived as encompassing six academic disciplines: linguistics, computer science, psychology, neuroscience, anthropology, and philosophy (Miller, 2003, p. 143). Now, roughly forty years later, the interdisciplinary endeavors are flourishing more than ever. Upon consideration of these endeavors, it becomes salient that the influence exerted by the various empirical sciences on philosophy of mind is highly diverse. Similarly, the influence of philosophy of mind on empirical sciences has many facets.

While each of the five papers in this dissertation targets a specific issue narrowly belonging to a particular area of discussion, they are all situated at the intersection between philosophy of mind and empirical sciences. In the following chapters I will – from a meta-perspective – show different ways in which one may approach this interdisciplinary field of study. Thus, the main purpose of the chapters in this introduction is to situate the papers of this dissertation in a broader context and show how they can be deployed *post hoc* to map and assess different kinds of interdisciplinary practice.

In section 1.1, I will preface a central caveat to interdisciplinary ventures involving philosophy of mind and empirical sciences. This caveat concerns the apparent contradiction in the possibility of reciprocal influence between *a priori* and *a posteriori*

academic practices. After introducing this caveat, I will (in section 1.2) provide a brief introduction to the broader interdisciplinary context in which the meta-perspective will aim to situate the papers.

In chapter 2, I present the academic background and the practices this dissertation belongs to. To this effect I will deploy a selection of interdisciplinary cases to exemplify the different ways in which philosophy of mind and empirical sciences may influence each other. Chapter 3 provides a discussion of methodology. In chapter 4, I summarize the papers. Finally, in chapter 5, I present the conclusions reached and situate them in the meta-perspective presented in chapter 2 in order to sketch possible directions for future research.

## 1.1 The analytic and the empirical

One pragmatic issue facing those involved in interdisciplinary endeavors is the task of reconciling the different methodologies, in the context of this dissertation these methodologies are those used in philosophy of mind and empirical sciences. While the methodologies (e.g. conceptual analyses, logical analysis, and thought experiments, to name a few) are relatively uniform across the different types of subject matter in philosophy of mind, this is not the case in empirical sciences. Even within a single subject area such as brain imaging, the applied methodology can vary significantly from one experimental paradigm to another. Furthermore, because many different areas of empirical science may be of interest to a given domain of philosophy of mind, and the methodology of each of these areas of empirical science is different from those of the others, the way in which each area meshes with philosophy of mind must be considered separately. This means that the task of applying empirical data to inform philosophy of mind has to be done carefully, on a case by case basis.

Importantly, the difference in methodological practice between philosophy and empirical sciences reflects a more fundamental difference between the academic practices. Philosophical argument traditionally rooted in conceptual analysis and logic resides squarely in the domain of *a priori* reasoning. This entails that, insofar as the one gets the premises and inferential steps right, the conclusion follows logically. In contrast to this, empirical sciences are, at their core, nothing but a collection of examples. Each of these examples is based on the observation of a phenomenon under some specific circumstances. This means that reasoning within empirical sciences is essentially inductive, i.e. *a posteriori* and contingent. This fundamental difference between philosophy and empirical sciences means that the quote by Miller prefacing

this chapter strictly speaking does not apply to the interdisciplinary ventures under consideration in this dissertation. Philosophy is held to yield *a priori* knowledge, established independently of empirical facts, while empirical knowledge is *a posteriori* and so depends on matters of fact. The reason why philosophical problems cannot be solved on empirical grounds is that philosophy concerns general knowledge and matters of principle that are independent of the actual state of the world and remain the same whatever happens. This epistemic issue is nothing new, and while it is worth bearing in mind, should not discourage us from pursuing knowledge in any way we can, including interdisciplinary ventures. However, we need to keep in mind that this difference in epistemic domain is of principal importance when considering the way in which the different domains may influence each other. Therefore, this issue will loom large throughout the discussions of interdisciplinary interactions in chapter 2.

## 1.2 The empirical turn in philosophy of mind

I think it is beyond doubt that the philosophy of mind has benefitted greatly from its early inclusion, in part thanks to Miller, in the interdisciplinary venture that became cognitive science. In parallel with – and most likely partly because of – its involvement in the cognitive science revolution, philosophy of mind has undergone an *empirical turn* of its own in recent decades. Characteristic of the empirical turn is an increased appreciation of the perceived benefit of complementing the philosophical enterprise with empirical data. Some philosophers (e.g. Weisberg, 2013) exhibit a great deal of optimism about the empirical turn and have gone as far as to suggest that the right way to approach some areas of philosophy of mind (e.g. consciousness) is through empirical data. In contrast to this optimism, it has been argued (Jackson, 1982, 1986; Levine, 1983; Nagel, 1974) that, at least with respect to consciousness, there is an explanatory gap between the objective empirical sciences and the subjective phenomena. More recently, it has been argued (Kriegel, *forthcoming*) that an explanation of the apparent correlation between consciousness and the brain may be, in principle, empirically underdetermined. A separate potential problem has been raised by McGinn (McGinn, 1991), who argues that certain features of the phenomenon of consciousness prevent a full explanation *simpliciter*. Be that as it may, it is worth noting that pessimistic views on the prospects of solving specific philosophical problems, such as those advanced by Levine, Jackson, Nagel and McGinn, do not seem to have any bearing on the overarching interdisciplinary enterprise of the empirical turn. There is no contradiction in holding the view that specific philosophical problems (e.g. subjective phenomena, according to Levine)

resist elucidation from empirical data, or that concrete phenomena (e.g. consciousness, according to McGinn) are beyond our explanatory capabilities in general, and still embracing the practices and advantageous prospects of the empirical turn.

On a final note, it is worth briefly mentioning the concept of *empirically informed philosophy*, a label that has gained some use as of late. Empirically informed philosophy is usually a label adopted by philosophers who are attuned to empirical developments pertaining to their field of study. Now, one might think that the concept of the empirical turn is co-extensive with empirically informed philosophy. However, because this is not the case, some clarification is needed to distinguish the former from the latter. As indicated by its name, empirically informed philosophy is engaged in a unilateral relation with empirical sciences, whereby the former receives input from the latter. Crucially, the empirical turn, as conceived of in this dissertation, involves more than that. It involves a beneficial reciprocity between philosophy of mind and the empirical sciences. The empirical turn is a concerted interdisciplinary venture, in which it is not only philosophers who invoke empirical data, but researchers in empirical sciences who utilize the capacities available to them in philosophers. Thus, because the empirical turn encompasses empirically informed philosophy *and* philosophically informed empirical science, there is a sense in which it subsumes the area of empirically informed philosophy.

# Chapter 2 - Background

The theme of this dissertation is to explore issues in the philosophy of mind where philosophical, conceptual theory intersects with empirical research. The five papers exemplify five (leaving it open that there may be more) major (and different) ways of approaching such issues.

This theme implies that philosophy of mind and certain areas of empirical science intersect with respect to the object of study. Much work that fits this theme has already been carried out by others, in what I in the introduction have called the empirical turn.

In this chapter, as a background to the papers, I provide a range of examples where philosophical, conceptual theory intersects with empirical research. I will begin by carving out two parameters useful for categorizing these examples. The two parameters map the strength and direction of influence between philosophy of mind and empirical science in a given interdisciplinary interaction. This mapping is useful as a rough sketch of the different roles the fields of philosophy and various empirical sciences may take. Additionally, the discussion in this chapter will serve to highlight some of the challenges and pitfalls facing interdisciplinary endeavors involving philosophy of mind and empirical sciences.

## 2.1 Directions and kinds of influence

The central assumption for the so-called empirical turn in philosophy of mind is that its area of research intersects with domains of empirical science. In order to obtain a better grip on the different kinds of interdisciplinary exchange between philosophy of mind and empirical sciences, two parameters that can be deployed to categorize the interdisciplinary exchanges comprise a useful analytic tool.

The first parameter, useful in mapping an interdisciplinary endeavor between philosophy of mind and empirical science, concerns the direction of influence. Call this the *dimension of application.* The dimension of application indicates whether a

given field of (or research program in, or experimental paradigm in) empirical science is applied to philosophy, or the other way around.

The second parameter tracks the amount of influence exerted by philosophy and empirical sciences upon each other. Call this the *dimension of impact*. The dimension of impact maps the strength of the influence exerted between the fields. Mild influence (e.g. work in one field informs existing work) at one end of the dimension of impact contraposes highly significant influence (e.g. one field inspires genuinely novel developments within the other) at the other end.

The dimension of application and the dimension of impact combine into a mapping system that we can deploy to characterize the interdisciplinary work of the empirical turn. Importantly, the two dimensions cannot – and are not intended to – provide an unequivocal categorization. In everyday practices the interdisciplinary interaction between philosophers and empirical scientists is a dynamic and ongoing affair. Nevertheless, as long as one bears this in mind, the parameters provide a useful analytic tool for mapping the interdisciplinary exchanges between philosophy of mind and related empirical sciences.

In the rest of this chapter, I consider examples of work that fall within the scope of the empirical turn. I will separate the examples on the basis of the dimension of application. In section 2.2, I will consider examples in which empirical sciences are applied to philosophy of mind. After that, in section 2.3, I turn to examples where philosophy of mind is applied to empirical sciences. The examples in each section are grouped according to the way the first academic field (i.e. empirical science in the first category and philosophy of mind in the second) exerts its influence on the second. Thus, the examples are not ordered according to the dimension of impact. Therefore, the fact that one example appears before another is not an indication that one or the other has exerted greater influence, or is more important or prevalent. However, at the end of the two sections I will discuss some high impact cases that one might consider belonging at the far end of the dimension of impact, owing to the novelty introduced by them. In my presentation of the examples, I will highlight potential caveats relating to the interdisciplinary application.

## 2.2 Empirical sciences influence philosophy of mind

This category describes practices in which empirical research influences philosophy of mind. I will start by considering examples located toward the mild end of the dimension of impact.

Many examples within this category and toward the mild end of the dimension of impact invoke empirical research in arguing for or against philosophical claims. Initially, one can divide these examples into three different practices. The first practice consists in cases where empirical research is proposed *in support of* a philosophical claim. This practice is responsible for the majority of work, both within this category and more broadly within the empirical turn. The apparent reason for the relative dominance of this practice is that, in general, researchers (understandably) are mainly concerned with the particular theories they espouse. Accordingly, the majority of work is allocated to developing and finding support for these theories. The second practice consists in cases where empirical research is *leveraged against* a philosophical claim. The third practice consists in *reassessment* of whether empirical evidence proposed in the first and second practices can do the work it is purported to do.

The kind of empirical research deployed in the three practices depends largely on the particular area of philosophy of mind under consideration. The scope of the empirical research and the philosophical claims ranges from highly domain specific to very generalized.

### 2.2.1 Empirical research proposed in support of a philosophical hypothesis

Starting with the first practice, in which empirical research is proposed in support of a philosophical theory or claim, I will outline three general approaches within this practice.

#### *2.2.1.1 Inference to the best explanation*

Classically, inference to the best explanation has been concerned with the move from evidence to the hypothesis that best explains it. As an early formulation by Harman (1965, p. 324) states:

> In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference.

This of course leaves open two questions. The first is: How do we determine whether a hypothesis best explains the evidence? The second is: What if we have a case where we cannot reject a competing hypothesis? Harman (1965, p. 324) addresses the first question when he continues:

> There is, of course, a problem about how one is to judge that one hypothesis is sufficiently better than another hypothesis. Presumably such a judgment will be based on considerations such as which hypothesis is simpler, which is more plausible, which explains more, which is less ad hoc, and so forth

The second question does not appear to have a ubiquitous answer. One possible answer may be that, in a case where we cannot decide which of two (or more) competing hypotheses best explains the evidence, further work is needed to resolve the issue. One way in which the second question is relevant to the current theme is, as I observe in the fourth paper (Kirkeby-Hinrup, 2016, p. 39), that in some cases it appears that competing theories are coherent and equally plausible from a purely conceptual point of view, and that the debate has reached a stalemate. In such a case one option is to look to the empirical data to help resolve the debate. If one pursues this option, the upshot is a two-tier process where inference to the best explanation of a piece of empirical data is part of a larger inference to the best explanation, where the latter aims to decide between competing theories that are equivalent from a conceptual point of view. This seems to be the version of inference to the best explanation Ned Block has in mind and advocates in his seminal 2007 paper. He writes (p. 486):

> I have in mind […] the familiar default 'method' of inference to the best explanation, that is, the approach of looking for the framework that makes the most sense of all the data […]

Ned Block (e.g. 2007, 2014) argues that results of the Sperling paradigm (Sperling, 1960) support his hypothesis that we consciously experience more than we can cognitively access (but see e.g. Schlicht, 2012 for an alternative interpretation, see also D'Aloisio-Montilla, *forthcoming* for a new empirical argument in favor of Block's hypothesis). In the classic Sperling paradigm, subjects were briefly presented with three rows of four letters and subsequently asked to report as many as they could. Despite stating that they saw all the letters, subjects would on average only be able to report three or four. The experimental manipulation demonstrated that, when cued to a specific set of the letters (i.e. one of the rows), the subjects could typically report every letter in the set. Given that the subjects did not know in advance which subset they would be cued to report, this supports the hypothesis that the subjects did experience all the letters, but subsequently could only access a limited amount. Block argues that when subjects report experiencing all the letters but can only report a limited subset this suggests the need for his conceptual distinction between phenomenal consciousness and access consciousness. According to Block, the fact that subjects can accurately report any of the rows, when cued specifically to one, gives

reason to believe that they are phenomenally conscious of all of the letters. Similarly, their inability to report all the letters suggests they can only be access conscious of a limited subset of that experience.

This way of using empirical research to argue for a hypothesis in philosophy of mind comes with significant caveats that one must keep in mind when assessing the significance. First, it is important to highlight that this way of reasoning is neither deductive nor inductive. Rather, the argument Block proposes in support of his distinction between access and phenomenal consciousness relies on abductive reasoning. Abductive reasoning is well illustrated in the original words of C. S. Peirce (1974, p. 137):

> Abduction makes its start from facts, without, at the outset, having a particular theory in view, though it is motivated by a feeling that a theory is needed to explain the surprising facts […]. In Abduction the consideration of the facts suggests the hypothesis.

Thus, similarly to inference to the best explanation (this similarity is also noticed by Harman, 1965), abductive reasoning can be characterized as proceeding from an observation to suggesting a hypothesis that may explain the observation (but see e.g. Frankfurt, 1958 for discussion of Peirce's concept of abduction). This returns us to the first question we can ask when faced with this way of applying empirical data to philosophy, which is whether the philosophical claim the empirical data is purported to support is the best explanation of the observation.

Second, there are two important differences between the traditional formulation of abductive reasoning given above and the application of abductive reasoning in many cases of philosophical work within the empirical turn (for a selection of cases see e.g. Block, 1995, 2007, 2008; Lau & Brown, *Forthcoming;* Lau & Rosenthal, 2011).

The first difference is that the reasoning often does not begin with an observation. Rather, the starting point is usually a philosophical hypothesis. This kind of reverse abduction, as it were, is indicative of a two-tier inference to the best explanation argument mentioned above, and should be treated with caution. The reason we should treat this kind of reverse abduction with caution is an inherent risk of confirmation bias, i.e., favoring evidence that supports one's pre-existing hypothesis.

The second difference is similar to the first, insofar as it concerns the kind of explanation proposed for the observational data. Presumably, the best explanation of an empirical datum is the one drawn by the researchers conducting the actual

empirical experiment. The explanations provided by the empirical researcher tend to be more frugal when it comes to implications of the empirical data, whereas philosophy by its very nature is mainly interested in drawing wider theoretical implications.

While they are worth being aware of, one should not be surprised or put off by the two differences between the normal understanding of abductive reasoning and its application within the empirical turn. For each of the two differences there is a caveat with regard to the abductive practices involved in the empirical turn. With regard to the first difference, the use of abductive arguments has the purpose of showing that the hypothesis in question is consistent with relevant empirical research; i.e., the hypothesis in question can account for a wide range of empirical phenomena. The motivation behind this way of deploying empirical data is to provide the basis for a (two-tier) inference to the best explanation of the kind Block suggests in the above citation. By showing that a philosophical hypothesis can account for a wide range of empirical data, the suggestion is that this hypothesis is the best explanation *overall*. This gives rise to the caveat that the philosophical hypothesis does not necessarily purport to be the best explanation within the narrow scope of specific empirical results for which it offers an explanation. Similarly, as alluded to above, the caveat with respect to the second difference is that philosophical theories are generally of a broader scope than the kind of strictly delimited and controlled settings we normally find in empirical research. This difference seems to pertain to interdisciplinary ventures in general, at least if one thinks that difference in domain implies difference in scope, and that the concept of interdisciplinarity suggests different domains.

#### *2.2.1.2 Empirical predictions*

Turning to an example of another kind of way in which empirical data may influence philosophy of mind, O'Regan, Myin and Noë (2005) suggest the study of sensory substitution (e.g. Sampaio, Maris, & Bach-y-Rita, 2001) as a promising avenue of research to test predictions of their sensorimotor theory of phenomenality. In sensory substitution the characteristics of input to one sensory modality (e.g. sight) are attempted to be translated into input to another sensory modality (e.g. tactile sensation). This research is especially pertinent to individuals who do not have access to one or more sense(s). In developing their sensorimotor theory O'Regan *et al.* proceed from the observation that the experience of, for example, driving a Porsche seems to depend on how the individual expects the surrounding context (i.e. the Porsche) to behave in response to her actions. These expectations of behavior are vetted by the individual through the sensory-motor loop when engaging in actions to manipulate the relevant surrounding context (the Porsche). When I press the

accelerator the car hums and speeds up. After doing this once, I know what to expect if I decide to press the accelerator again, and this expectation becomes nested in my Porsche-driving experience. In this way, O'Regan *et al.* argue, the acquisition of a set of expectations between motor output and sensory feedback is central to phenomenality. It consists in the implicit knowledge of how the sensory input would change in response to an action of the individual (O'Regan *et al.*, 2005, p. 371). The prediction by O'Regan *et al.* is that the success of sensory substitution devices will rely on the output of the device, and specifically the output's ability to imitate the sensorimotor laws of the modality to be substituted. The authors (O'Regan *et al.*, 2005, p. 381) write:

> …it will be the similarity in the sensorimotor laws that such devices recreate which determines the degree to which users will really feel they are receiving stimulation in the modality being substituted.

The sensorimotor laws are presumed to be characterized by modality-specific processing and shared among individuals, which makes them amenable to testing.

A good example to illustrate the authors' notion of sensorimotor laws (that is not brought up by the authors themselves), is the Doppler effect in auditory sensation. The Doppler effect is the change in perceived frequency as a source of sound moves past an observer, such as the change in perceived frequency of a passing police siren. The issue then would be the possibility of constructing a sensory substitution device that can successfully replicate the Doppler effect in another sensory modality than the auditory (e.g. in the visual system).

O'Regan *et al.* (p. 381) propose sensory substitution as an opportunity to test a (according to them) counter-intuitive prediction in that:

> It should be possible to obtain a visual feel from auditory or tactile input, for example, provided the sensorimotor laws that are being obeyed are the laws of vision (and provided the brain has the computing resources to extract those laws).

By providing an (in principle) empirically testable prediction, the authors can be seen to perpetuate interdisciplinary interaction. Clearly, should the empirical results conform to the prediction provided by the proponents of the sensorimotor theory, this would be a good example of fruitful interdisciplinary interactions in the empirical turn.

One general point worth noting is that, when providing testable empirical predictions, one should be sensitive to the results, whether they confirm or disconfirm one's prediction. Echoing Popper, upon hearing a hypothesis of mine and my

examples of different empirical results it could explain, a prominent empirical researcher replied: "What would it take to disprove your idea?"

Returning to the sensorimotor theory, there is a sense in which one might suggest that owing to its empirically minded character it is better classified as an empirical theory rather than a philosophical one. If this is the case, the sensorimotor account would be in the business of theory development in the empirical domain. It is, however, worth noting that the same considerations apply if we categorize the sensorimotor theory as an empirical – rather than philosophical – theory. In this case, even if the empirical predictions fail to be confirmed, this would not necessarily falsify the theory. Given that the theory is more general in scope than the specific predictions and moreover is concerned with fundamental theoretical questions pertaining to its domain, it may need revision in the face of unexpected empirical data, but not outright disbandment. Conversely, if the predictions are confirmed this would constitute significant support for the theory.

#### *2.2.1.3 Conceptual mapping*

Zahavi and Rochat (2015) deploy empirical data in yet another way. Zahavi and Rochat present a minimalist theory empathy that draws on insights from phenomenology and data in developmental psychology. According to Zahavi and Rochat, empathy is a basic sensitivity to the mindedness of others and involves neither metacognitive processes, simulation of the other's mental states, or a fusion of perspectives. Most importantly, it does not involve affective sharing. Zahavi and Rochat suggest that empathy serves as the foundation for three stages of ontogenetic development of 'we-ness' identified in developmental psychology, and that it might be a precondition for sharing. One conclusion Zahavi and Rochat draw from the minimalist theory of empathy is that the notion of emotional sharing involved in many discussions of empathy conflates sharing with similarity and does not recognize that reciprocity is involved in sharing proper (Zahavi & Rochat, 2015, p. 543). Zahavi and Rochat suggest that careful psychological observations are useful for the development of a more sophisticated concept of emotional sharing. In order to illustrate their view that sharing involves reciprocity, Zahavi and Rochat point to the notion of joint attention. In order for two individuals jointly to attend a scene, object or event, the attending must not only merely be parallel. Rather, it must involve an awareness of attending together; the fact that the two individuals are attending the same scene, object or event must be mutually manifest. As an example of the separation of sharing and empathy set up by Zahavi and Rochat by reference to developmental psychology consider the following quote:

> […] Given the minimalist definition of empathy provided by the phenomenologists, where empathy rather than being identified with, say, prosocial behavior or a very special kind of imaginative perspective taking, is simply used as a label for our most basic other-acquaintance, i.e., our sensitivity to and direct experience of other minded creatures, it should be fairly obvious that empathy is presupposed by all the early dyadic and triadic types of sharing.
>
> Zahavi & Rochat, (2015, p. 551)

The upshot is a separation of the notions of empathy and sharing, where the former does not entail the latter, but may be a precondition for it. This conclusion runs counter to what the authors characterize as a widespread view of empathy, in which empathy is viewed as the process in which an individual comes to have identical affective experiences to another (Zahavi & Rochat, 2015, p. 551).

The parallels drawn to developmental psychology support the minimalist theory of empathy empirically by illuminating the philosophical claims and showing that there are empirical counterparts to these. Similarly, the philosophical claims may serve to provide conceptual clarifications of the empirical data. This kind of theoretical bridge building between philosophical theory and empirical findings allows the former to illuminate the latter and *vice versa*. Furthermore, the conceptual mapping between the minimalist theory of empathy and the empirical findings from developmental psychology provide a foundation for drawing and discussing hypothetical consequences.

### 2.2.2 Leveraging empirical research against a philosophical hypothesis

The second practice is leveraging empirical evidence against a philosophical claim. Arguing against competing theories has been mainstay philosophical practice since the Greeks. From the perspective of the empirical turn, deploying empirical evidence to argue that a philosophical claim is empirically implausible can be seen as an extension of this practice. Importantly, when engaging in or considering the implications of this practice, one needs to remember the distinction between empirical evidence and conceptual proof. To recapitulate this distinction is useful as a precursor to evaluating the examples below. When invoking empirical data as evidence against a philosophical claim, we cannot purport to show that the philosophical claim, insofar as this is a conceptual matter, is false *per se*. The implications of leveraging empirical evidence against a philosophical claim cannot extend beyond showing that the claim in question is empirically implausible, i.e. that it is not corroborated by the received

view within a relevant field of empirical investigation. This, in turn, amounts to presenting an (extra-theoretical) motivation for reconsidering or revising the claim, and/or the philosophical theory in the context of which it is advanced. The fundamental difference in kind between philosophical/conceptual argument and empirical observation entails that empirical evidence can, at best, provide indirect (i.e. extra-theoretical) evidence against a philosophical/conceptual claim. To embrace any stronger implication of empirical evidence on conceptual matters is tantamount to committing a naturalistic fallacy (for an early assessment of the use of this term, see e.g. Frankena, 1939). Traditionally, the naturalistic fallacy states that ethical conclusions are not warranted from premises that are not ethical, i.e. that ethical propositions cannot be deduced from non-ethical ones. The central idea, and the reason I draw the parallel to the empirical–conceptual issue, is that the naturalistic fallacy concerns the validity of inferences between domains that are epistemologically insulated from each other (e.g. what is observed, on the one hand, and what is *a priori*, on the other).

With this caveat acknowledged, it is nevertheless good scientific practice to evaluate connections between philosophical claims and empirical data, as long as one bears the described caveat in mind. Evaluating connections between philosophical claims and empirical data is especially relevant with respect to philosophical theories of mind that aspire to the project of naturalizing the mind. The kind of naturalization I have in mind here is akin to the one proposed by Jean Petitot *et al.* when they write: "By 'naturalized' we mean integrated into an explanatory framework where every acceptable property is made continuous with the properties admitted by the natural sciences" (Petitot, Varela, Pachoud, & Roy, 1999, pp. 1–2). This way of conceiving naturalization seems to come with a commitment to empirical corroboration of philosophical theory.

One philosophical theory with naturalistic aspirations is the higher-order thought theory of consciousness advocated by David Rosenthal among others. In his (2014) paper, Miguel Ángel Sebastián leverages brain imaging results from sleeping subjects against the higher-order thought theory of consciousness. Specifically, Sebastián targets the hypothesis that the dorsolateral prefrontal cortex is involved in the production of (consciousness generating) higher-order thought. Sebastián argues that because brain imaging during dreaming shows reduced activity of the dorsolateral prefrontal cortex, theories of consciousness (e.g. Lau & Rosenthal, 2011) that claim this cortical region has a role to play in phenomenal consciousness are implausible. The reason is that dreams arguably have some phenomenal properties. We should therefore expect very little decrease in activation in any brain region conjectured to underpin phenomenal experiences. The appreciation for the way of arguing indirectly

against philosophical theories via empirical data exemplified by the second practice is on the rise among empirically minded philosophers. This is evident in Weisberg's (2013, p. 433), a prominent proponent of the higher-order theory of consciousness, reply to Sebastián:

> I think Sebastián presents a strong challenge to the theory and a challenge, refreshingly, from the empirical side of the road. This is the proper way to approach consciousness, rather than taking endless detours to zombie worlds and the color deprived prisons of super scientists.

In the previous example Sebastián leveraged a concrete empirical datum against a specific theory of phenomenal consciousness. An example of empirical data being leveraged more broadly against a group of theories is found in the results from Benjamin Libet (2004). Libet's results have been widely deployed to suggest that theories of free will involving a conscious component efficacious in initiating actions are empirically implausible. Libet used an electroencephalogram (EEG) to measure the readiness potentials in the motor cortex and compared these to the timing of the subject's decision to act. Libet's results gained much attention among philosophers working in the domains of free will and voluntary action, because they showed that movement had already been initiated in the motor cortex as much as 300 milliseconds before the individual (reported that she) decides to act. The upshot of the empirical data is the suggestion that the conscious decision to act is epiphenomenal and plays no role in actually initiating an action. Now, as reiterated above, we are not at liberty to claim that *because* of the empirical data theories of free will relying on a conscious component are *proven* to be false. The most Libet's experiment warrants is, on the one hand, a motivation to reconsider and possibly revise philosophical theories in light of the empirical findings and, on the other hand, to investigate further the avenue of empirical research pioneered by Libet to corroborate the results.

### 2.2.3 Re-assessing arguments from and interpretations of empirical evidence

The third practice is to evaluate whether the interpretation of empirical research and the arguments derived from it in support of (or against) a particular philosophical theory or claim can withstand scrutiny. Thus, the third practice consists in scrutinizing and reassessing the arguments and interpretations proposed by the former two practices, such as the effort of Tobias Schlicht (2012) in reassessing the arguments of Ned Block mentioned above. This makes the third practice the philosophical equivalent of replication experiments in empirical sciences. As with replication experiments in empirical science, this practice is underappreciated and has

received relatively little attention. However, I submit that this practice deserves more attention because it serves as a safeguard against the potential pitfalls involved in the first two practices as discussed above. The way in which it might serve as a safeguard is by scrutinizing the steps of abductive reasoning involved in the first two practices. There are at least two basic and fairly reliable ways to go about scrutinizing such abductive reasoning and re-assess the way in which empirical results are brought to bear upon a philosophical claim. The first method is to assess whether the suggested interpretation of the empirical results is reasonable, or whether there are alternative interpretations that are equally (or more) reasonable. The second method is to assess the connections between a given interpretation of the empirical data and a philosophical claim. This will usually mean investigating how the interpretation features in an argument purporting to arrive at a particular conclusion in the context of a given philosophical claim.

One example of the first method is given by Rafael Malach (2011), who argues that one instance of the empirical evidence suggested by Hakwan Lau and David Rosenthal (2011) in support of the higher-order theory of consciousness, when subjected to a competing interpretation, in fact contradicts the theory. The empirical evidence in question is derived from Goldberg *et al.* (2006). What Goldberg *et al.* set out to investigate was whether self-related processes were necessarily engaged in sensory perception. They investigated this using functional magnetic resonance imaging (fMRI), comparing activity patterns involved in a demanding sensory categorization task with those engaged in an introspective task. The main result from the study was a complete separation of self-related cortical regions active in the introspective task (e.g. the prefrontal cortex) and the sensory-motor regions involved in the categorization task. Additionally, the self-related regions were inhibited during the categorization task (Goldberg *et al.*, 2006, p. 336). The authors hypothesize that the reduced self-related activity indicates that self-related processes are not necessary for subjective awareness (Goldberg *et al.*, 2006, p. 337). Lau and Rosenthal argue that the results from Goldberg *et al.* showing reduced activity in the prefrontal cortex while the subjects experienced the visual stimuli with degraded detail corresponds well with what their theory would predict. They do this on the basis of the idea that the prefrontal cortex is involved in the generation of conscious thought, because of its role in generating higher-order thoughts. Malach (2011) targets this interpretation of Goldberg *et al.* (2006) by arguing that the general conclusion from Goldberg *et al.* is that the reduced prefrontal activation actually is a result of suppression: "…during intense perceptual engagement, all neuronal resources are focused on sensory cortex, and the distracting self-related cortex is inactive" (Goldberg *et al.*, 2006, p. 337). If this is correct, then *pace* the interpretation suggested by Lau and Rosenthal, the

prefrontal cortex is not critically involved in conscious perception. What Malach does in his paper, and what makes this a good example of the third practice, is to show that there is a competing interpretation of the empirical findings of Goldberg *et al.* to the one suggested by Rosenthal and Lau. In addition, Malach shows that this competing interpretation actually contradicts the theory the suggested interpretation was supposed to support. It is worth noting that the difference in interpretation of the empirical data highlighted by Malach is an instance of the caveat mentioned in section 2.2.1.1 where philosophers diverge from the (generally more frugal) interpretation of empirical data proposed by the empirical scientist responsible for collecting the data in question. Given that the interpretation promoted by Rosenthal and Lau diverges from the interpretation offered by Golberg *et al.* there is reason to subject it to additional scrutiny, to validate the inferences upon which the interpretation rests.

### 2.2.4 High impact cases

Moving up along the dimension of impact, I will now provide a selection of cases where empirical research has had a significant impact on philosophy of mind. For the most part, these are examples of empirical research prompting new developments in philosophy of mind. Generally, to be considered as belonging to the far end of the dimension of impact, the development within philosophy of mind prompted by empirical research should conceivably involve something relatively *new*. Examples of such new developments may consist in new theories within a given subject area, or significant variations of existing theories. A hypothetical example could be a new explanation of a key component in a theory, such as a new way to conceive of the mechanism responsible for feeding unconscious mental states into consciousness. Of course, criteria for exactly when something is to be counted as *new* are likely to involve a certain amount of arbitrariness. However, because the examples are conceptualized along a dimension, there is no exact cut off separating development of something new and merely significantly influencing an existing theory. What matters in this section is highlighting some examples that one might say have provided new developments.

In the above, the results of Benjamin Libet regarding the timing of voluntary action were discussed as an example of how empirical data influenced philosophy of mind. The upshot of the previous discussion of Libet's results above was that the common interpretation appeared to tell against theories of free will that invoked a conscious component. In stark opposition to this common interpretation, Benjamin Libet (2004) advances his results as the basis of a novel theory of free will. As explicated

above, Libet found that signals from the motor cortex initiating movement preceded conscious experience of intention to move by several hundred milliseconds. However, Libet points out, there is a window of around 150 milliseconds between the conscious intention and the movement actually being executed. Libet suggests that this window potentially allows the conscious intention to influence the movement, *to wit* by halting it. This leads to what can be dubbed the *veto theory* of free will (see e.g. Bonn, 2013; Gomes, 1999 for theories that draw on this idea). Being a novel philosophical theory spurred by empirical research, the veto theory of free will can be viewed as belonging in this section as a case of empirical data having a high impact on philosophy of mind.

Another example of empirical research inspiring an entirely new theory is the split-brain syndrome (SBS), which has instigated the discussion of several philosophical theories concerning the unity of consciousness and the concept of personhood. One such theory is the *switch* theory of Tim Bayne (2008). The SBS was originally discovered in subjects who had undergone a callosotomy. In a callosotomy, the corpus callosum (the major nerve bundle that connects the right and left brain hemispheres) is severed, usually to relieve the effects of serious epileptic seizures. In ordinary interactions, the procedure itself appears to have surprisingly few symptoms in the overt behavior of the subject. However, under controlled experimental conditions curious effects become manifest. Because most sensory input is contralateral, meaning that, for example, visual input to the right eye is processed in the left brain hemisphere, careful control of such input can yield surprising results. The curiousness of these results is underscored by the fact that generation of overt speech behavior is located in the left hemisphere. This means that a split-brain subject can only verbally report on a visual stimulus if this stimulus reaches the left brain hemisphere. Thus, when a visual stimulus of an item is presented exclusively to the left eye (whose input is propagated to the right hemisphere) and the subject is asked to pick up the item shown from a selection of items in front of her, the subject will verbally express bewilderment, with exclamations of the sort "you did not show me any item". All the while, the left hand picks up the relevant item and presents it to the experimenter. This effect is the result of the visual stimulus not being present in the speech-controlling left brain hemisphere, whereas the presence of the stimulus in the right hemisphere allows the subject to fulfill the experimenter's request using the contralaterally controlled left hand. The pressing question, given these results, is whether consciousness is phenomenally unified in the way it is normally presupposed to be (see e.g. Bayne & Chalmers, 2003 for considerations on this presupposition). In the face of this question Bayne developed the switch model to defend the possibility that consciousness is in fact still unified in split-brain cases. According to the switch

model proposed by Bayne, consciousness fluctuates rapidly between the left and right hemispheres. This means that, rather than having to provide an account that unifies the simultaneous processes in the right and left brain hemispheres, the switch model states that phenomenal consciousness is present in only one hemisphere at a given time, and switches back and forth between the hemispheres. This means that at any given point in time consciousness will be phenomenally unified, in the sense that the occurring phenomenal states in the currently conscious hemisphere are synthesized into a whole (see e.g. Bayne & Chalmers, 2003, p. 12).

Turning to an example where empirical research does not prompt an entirely new theory, but rather a significant variation of an existing philosophical theory, let us consider the higher-order Bayesian decision theory proposed by Hakwan Lau (2007). Lau presents a formal theory of perceptual consciousness based on signal detection theory (SDT). Using forced choice visual detection tasks as an example, Lau suggests that the strength of the visual signal can be interpreted as a probability distribution. The probability distribution is matched against a subjective decision criterion to evaluate whether to answer "yes" or "no" in the forced choice detection task. Lau argues that interpreting internal signals (e.g. visual signals from visual stimuli) as probability distributions is necessary because the brain is essentially a *noisy* detection system. The variability of signal strength is in part a consequence of this noisy environment, and to this end operating on probability distributions allows the detection mechanism a dynamic way to determine whether an internal signal actually carries information or is just noise. Lau frames his higher-order Bayesian decision theory as a variant of the higher-order thought theories of consciousness. This allows Lau to deploy the SDT model to describe the mechanism that provides input for higher-order representation. Furthermore, because interpreting signals according to probability distributions comes with an inherent possibility of false-positives and false-negatives, the SDT model can account for abnormal cases of perceptual consciousness, e.g. blindsight (Weiskrantz, 1986).

Another example of empirical discoveries prompting significant developments to a philosophical theory can be found in the domain of theory of mind, i.e. the ability to adopt (cognitively or emotionally) the perspective of others. The discovery of mirror neurons in macaque monkeys (Di Pellegrino *et al.*, 1992) ushered in a new era for the so-called simulation theory. The simulation theory proposes that the way in which individuals adopt the perspectives of others is by simulation, as opposed to by inference (the latter is advocated by the proponents of the so-called theory-theory). Mirror neurons were originally identified as clusters of neurons in the ventral premotor cortex F5 (they have later been found in other regions as well) that were found to be active both when individuals perform an action and when they observe

others perform the action. The latter kind of activation provided the early versions of simulation theory (e.g. Gordon, 1986) with empirical data that could support the idea that offline imitation (*viz.* simulation) of the behavior of others is an empirical corollary to the philosophical idea of simulation. On the basis of this discovery Vitorrio Gallese and colleagues (e.g. Gallese, 2007; Gallese & Goldman, 1998; Gallese & Sinigaglia, 2011) have continued to develop this new branch of simulation theory.

## 2.3 Category 2: Philosophy of mind influences empirical science

In this section, I will present examples where the philosophy of mind influences empirical research. One issue facing considerations of this category is that the influence of philosophy of mind upon the theories and practices of empirical sciences can often be subtle. This might lead one to think that the interdisciplinarity involved in the empirical turn is generally unilateral, going from empirical science to philosophy of mind. This, I think, is merely a matter of appearance. It appears, from, for example, conceptual clarifications and perspectives drawn, that a range of theories, paradigms, and publications in empirical fields are likewise influenced by philosophical work. The issue therefore might mainly be that this influence does not generally manifest itself in citations. There are however notable examples.

### 2.3.1 Conceptual clarification

In his seminal paper, David Chalmers (1995) formulates the distinction between the hard problem of consciousness and the so-called easy problems. The hard problem of consciousness poses the question why and how it is that some organisms are subjects of experience. Why and how is it that, when our cognitive systems engage in certain forms of information processing, this gives rise to experiential qualities such as the subjective feel of listening to Wagner or tasting a cherry? What Chalmers highlights in his paper is that, at the time, many empirical researchers who considered themselves to be working on consciousness were not addressing this central problem. Instead empirical research appeared to be focused mainly on investigating the easy problems of consciousness. The so-called easy problems include, for example, how a cognitive system integrates information, and what underlies the reportability of mental states and the ability to discriminate, categorize and react to stimuli.

Importantly, Chalmers is not out to diminish the challenge posed by – or the importance of – the easy problems, but rather to illuminate that many empirical approaches to consciousness are really addressing very specific areas, and research purporting to investigate consciousness *simpliciter* is usually overstating the case. The upshot is an appeal to stay frugal with regard to claims pertaining to empirical investigations of consciousness. The distinction between the hard and the easy problems of consciousness has won wide recognition and has become a mainstay among empirical researchers working on consciousness and cognition (Levy & Anderson, 2012; Tallon-Baudry, 2012; Wessel, 2012).

Another area in which philosophy has had an impact on empirical sciences is artificial intelligence (AI), where the *Chinese room* thought experiment by John Searle (1980) has won recognition as a difficult challenge. Searle argues that because AI is essentially following predefined rules for connecting input to output, there can never be support for the claim that it has humanlike qualities, such as intentionality and understanding. This poses a problem for the development of AI insofar as the aim is to produce something that is similar to human intelligence, what Searle dubs *strong AI*. According to Searle, strong AI is impossible if it is solely instantiated in a software algorithm, because a certain physical basis appears to be needed in order to have the ability to, for example, understand a language. Conversely, merely producing a *weak AI*, i.e. something that mimics human behavior, is achievable because such behavior can be produced by a sufficiently complex algorithm disregarding the physical realization of the function. Searle's considerations on strong and weak AI have received widespread attention as a central caveat in relation to the prospects of strong AI (e.g. Dowe, Hernández-Orallo, & Das, 2011; Mayo, 2003; Nilsson, 2005).

### 2.3.2 Methodological considerations

Ned Block is the source of another central distinction widely acknowledged among empirical researchers working on consciousness. In his classic paper "On a confusion about a function of consciousness" (Block, 1995), Block calls consciousness a mongrel concept (p. 227) in need of explication. In the paper, Block identifies two main functions that appear distinct, albeit intertwined. On one side of the distinction Block puts *access consciousness*. Access consciousness covers the cognitive machinery underpinning the ability to deploy information in inferences, action control and reports. On the other side of the distinction Block puts *phenomenal consciousness.* Phenomenal consciousness is essentially the experience of phenomenal qualities associated with the mental states of the individual. Importantly, access consciousness and phenomenal consciousness are intertwined and need careful disentanglement to

avoid confusion of empirical results. For instance, subjective reports of experiences seem to require that parts of phenomenal consciousness be reported, but this practice relies essentially on access consciousness, because without access the individual cannot report their experiences. This fosters a *methodological puzzle* (Block, 2007). The methodological puzzle occurs when a subject fails to report an experience we have reason to believe she has. The question then is whether she does not have the experience, i.e. the experience does not figure in the subject's phenomenal consciousness, or whether the subject has the experience but cannot report it because it is not available to access consciousness. The distinction between access consciousness and phenomenal consciousness has received widespread attention in empirical research (e.g. Fleming, Dolan, & Frith, 2012; Kouider, De Gardelle, Sackur, & Dupoux, 2010; Lamme, 2004). For instance, Victor Lamme (2003, 2004) explicitly applies the distinction between access consciousness and phenomenal consciousness in his investigations of conscious visual awareness. Lamme proposes that phenomenal consciousness (awareness) depends on recurrent feedback in the visual cortex, and demonstrates this by disrupting this recurrent feedback using successive visual stimuli, thus introducing competition for cognitive processing among the stimuli. According to Lamme, this competition explains why access consciousness is often conflated with phenomenal consciousness. The reason is that, in access consciousness, recurrent interactions of visual areas integrate with action- or memory-related areas, and awareness evolves from phenomenal to access awareness (Lamme, 2003, p. 16).

It is worth noting that, in addition to the distinction between access and phenomenal consciousness, the methodological puzzle as explicated by Block has influenced debates on the methodology concerning subjective reports in empirical paradigms (Brogaard, 2011; Hohwy, 2012; Overgaard, 2015; Overgaard & Mogensen, 2014).

### 2.3.3 Explicit input

The contemporary academic journals related to philosophy of mind contain a host of ideas and comments on empirical research from philosophers. Supposedly, this work has value, even if not all the efforts result in citations by – or collaborations with – empirical researchers. In order for the empirical turn to remain a truly bilateral interdisciplinary venture, philosophers cannot simply be satisfied with empirically informed philosophy. The burden is on the philosophers to maintain a flow of input to the empirical sciences.

One way to manifest this flow of input is by offering up explicit suggestions of future avenues of empirical investigation tied to specific questions within our field, perhaps even concrete experimental paradigms. One effort in this regard is the suggestion to investigate further the possibilities within sensory substitution in relation to the sensorimotor theory of consciousness discussed above in section 2.2.1.2.

Another effort in this regard is described by Brinck (2015), who suggests a way to improve an experimental paradigm investigating the development of the understanding of social norms in children (Rakoczy, Warneken, & Tomasello, 2008). The empirical paradigm involves a child interacting with two dolls in playing a game. Brinck argues that the experimental paradigm developed by Rakoczy *et al.* fails to test accurately for children's understanding of social norms, but instead captures their understanding of constitutive rules. Brinck proceeds to propose a re-interpretation of the existing data in the light of constitutive rules. Additionally, Brinck suggests a change in the experimental design by altering the interaction between one of the puppets and the child in the experimental paradigm, in order for the puppet's behavior to reflect the breaking of social norms involved in playing a game, as opposed to simply misunderstanding the rules of the game (Brinck, 2015, p. 713).

Another way to manifest this flow of input is to perpetuate the conceptual development pertaining to a given empirical area of research. Actively assisting in the interpretation of experimental data by offering conceptual clarifications of central and auxiliary facets of the data is a concrete way to facilitate the interdisciplinary approbation (and possibly alleviate growing pains in emerging fields) of empirical research and is related to developing the conceptual framework of existing theories to, for example, accommodate recalcitrant data (e.g. Brinck, 2001, 2004).

### 2.3.4 High impact cases

In this section, I will briefly mention examples where philosophy of mind has had a high impact on empirical science. As with the high impact cases mentioned in section 2.2.4, I here take high impact to imply some amount of novelty.

In exactly this vein, a new kind of philosophical enterprise has emerged in within the last two decades: experimental philosophy. Experimental philosophy applies methods most commonly associated with psychology and social sciences to carry out investigations to shed light on philosophical questions and debates. Most prominently, surveys have been carried out, mainly on non-philosophers, to map folk psychological conceptions of consciousness (Knobe & Prinz, 2008), moral intuitions (Nichols & Knobe, 2007), knowledge ascription (Weinberg, Nichols, & Stich, 2001)

and intentional action (Knobe, 2003b). The last is a useful example to illustrate the methodology and application of experimental philosophy.

Knobe's studies on intentional action are situated in a debate among competing theories as to what qualifies as intentional action. One theory suggests that the concept of intentional action is inherently tied to theory of mind, i.e. to the prediction and explanation of behavior. A competing theory suggests that intentional action is essentially tied to normative evaluations. According to this latter theory the concept of intentional action is only properly understood if we factor in its role in determining the moral significance of the action. In order to investigate these competing theories of intentional action, Joshua Knobe (2003a, 2003b) conducted surveys to elucidate folk psychological ascriptions of intentional action in hypothetical cases. In the surveys subjects were presented with vignettes describing different situations and were asked to judge whether the individual acting in the vignette was acting intentionally. Knobe argues that his findings show that normative evaluations bleed into the ascriptions of intentional action. Thus, Knobe argues, his research strongly supports a theory of intentional action in which the concept is deeply intertwined with normative evaluations.

While I am sympathetic to the project of experimental philosophy, it is worth noting that this field is subject to some of the same worries concerning the role of empirical data in philosophical debates discussed above. For instance, empirical data arising from the practice of experimental philosophy still cannot be applied directly to philosophical arguments without committing a corollary of the naturalistic fallacy. This means that empirical data obtained via experimental philosophy can at best serve as auxiliary evidence to philosophical claims, demonstrating that these have empirical counterparts or are empirically plausible with respect to the sampled population. Two other commonly raised objections to experimental philosophy are worth mentioning. The first is that experimental philosophy places too much significance on intuitions. This sentiment is reflected by Max Deutch when he says: "It almost never comes down to intuitions […] in philosophy, it all comes down to arguments" (Deutsch, 2010, p. 457). Similarly, it has been objected that the intuitions surveyed in many experimental philosophy cases are not the intuitions that are relevant (see e.g. discussion in Feltz, 2009). The underlying idea here is that while intuitions may be important in philosophy, we should rely mainly on so-called *expert* intuitions, i.e., the intuitions of individuals well acquainted with the subject in question.

An excellent example of, on the one hand, the methodological challenges and, on the other hand, the promise of experimental philosophy can be found in an article by Gunnar Björnsson (2016). On the basis of perceived shortcomings of an earlier experiment by Sripada (2012), Björnsson sets out to replicate the experiment with

alterations to control for the possible shortcomings. Sripada originally set out to investigate whether incompatibilist or compatibilist intuitions drive the ascription of free will and responsibility in a vignette describing a manipulated agent. The incompatibilists hold that intuitions in such cases are sensitive to a manipulated agent not having ultimate control over her actions. Conversely, compatibilists argue that the main factor in our intuition pertains to damage to the psychological and volitional capacities of the agent. Sripada concludes that when subjects judge that a manipulated agent was unfree, this is fully explained by their judgment that she had suffered damage to key psychological capacities. Sripada takes this to support the idea that intuitions on free will are driven by underlying compatibilist intuitions.

Björnsson (2016, pp. 640–644) raises a variety of worries in relation to Sripada's data. The most significant of these pertain to the wording used in the vignette. Thus, Björnsson carries out a replication experiment with minor changes to the wording in the vignette, to better assess the robustness of Sripada's findings. Surprisingly, Björnsson's findings show almost the opposite of Sripada's, i.e. that incompatibilist intuitions are driving the judgments that the agent was unfree. Thus, by making minor alterations to the vignette, Björnsson found that subjects deemed the agent to be unfree mainly owing to external factors outside of their control, as opposed to damage to psychological capacities. This stark contrast in the findings of two almost identical studies highlights the methodological challenges for experimental philosophy. Among these challenges is the fact that there is mounting evidence of significantly varying intuitions between individuals, determined by a variety of factors including personality, culture, cognitive style, religious commitment and socio-economic status (Feltz, 2009). This raises questions about whether results generalize, and highlights the impact of the selected subjects on the results. This means that scientific rigor is necessary in order to obtain broadly applicable and convincing results. The other (and bright) side is that the scientific rigor exemplified by both Sripada and Björnsson speaks to the promise of experimental philosophy. The possibility of performing – and willingness to perform – replication experiments on large numbers of subjects (the Sripada study had 240 subjects and Björnsson's had 361) is a significant strength of experimental philosophy. Moreover, the thorough analysis both authors apply to the data allows the development of statistical models that may serve as foundations for further research.

As an example of empirical research carried out explicitly in relation to a philosophical debate, Grush *et al.* (2015) carried out an experiment to investigate the age-old philosophical problem popularly termed the "inverted spectrum". Specifically, Grush *et al.* tested whether phenomenal color adaptation occurred when subjects wore LCD goggles attached to a camera, where the goggles displayed the color

spectrum rotated by 120 degrees (i.e. blue appeared green, green appeared red and red appeared blue). What Grush *et al.* were interested in testing was whether phenomenal adaptation would occur within the testing period. The group analyzed its results in light of three groups of theories about what phenomenal adaptation would imply. Thus, the question was whether, after a wearing the color inverting goggles for a while, the perception would adapt and return to "normal". In the context of the analysis of Grush *et al.,* the three groups of theories are distinguished by the proposed role of phenomenal properties. The first group of theories classically embraces phenomenal properties as experiential qualities (qualia). The second group of theories views phenomenal properties as the vehicles for enactive sensorimotor (and related) contingencies. The third group of theories advocates that phenomenal properties merely play a role in discriminatory tasks. The results of the research with color inversion, according to Grush *et al.*, provide the most support for the last of the three groups, while providing some evidence to counter the plausibility of the first.

In this chapter, I have considered a range of examples of interdisciplinary interactions between philosophy of mind and the empirical sciences. I have broadly categorized these examples into two categories depending on the direction of influence. Within each category, I have given examples of different practices depending on how philosophy of mind was applied to empirical science and *vice versa*. I have raised several methodological caveats related to various practices, as well as highlighting strengths and motivations. I submit that, when viewed as whole, the examples provide a diverse (but, owing to space constraints, limited) picture of the kind of ongoing interdisciplinary endeavors that make up the empirical turn.

# Chapter 3 - Methodology

The methods deployed vary among the papers in this dissertation, but all the papers draw on the traditional practices of analytic philosophy. Each paper addresses a specific question in relation to which I make one or more arguments. As a consequence, the question, and the argument(s) I make in relation to it, taken together constrain which methods can be relevantly applied.

The first paper takes its methodological starting point in careful scrutiny of the specific empirical data mentioned in relation to the philosophical hypothesis I am investigating. Because the empirical data are deployed in an argument to support a philosophical hypothesis, *to wit* by demonstrating putative empirical cases of higher-order misrepresentation, the next methodological step is to clarify the premises and conclusion of the argument in order to assess them. The method applied here is best characterized as argumentative analysis and reconstruction. The aim is to identify the reasoning in the literature, to illuminate the argument. In doing this I seek to adhere to the principle of charity wherever necessary, granting additional premises to the ones explicitly mentioned, in order to make the argument valid. In evaluating the argument I apply conceptual analysis and informal logic. This application leads to the identification of an ambiguity, the resolution of which yields the identification of a two-pronged dilemma.

In the second paper I take a more traditional philosophical approach. This is reflected in the methodology applied. I take my departure in a conceptual analysis of a central concept (free will) of a particular theory and argue that it has shortcomings in terms of how this concept is usually conceived. The identification of these shortcomings is based on comparison of the implications of the concept with prevalent intuitions. I diagnose the theory in question in order to determine the aspect that gives rise to the shortcomings. After identifying this aspect I proceed to develop an alternative mechanism to play the role previously handled by the problematic aspect. In developing this mechanism I deploy examples to illustrate the workings of the mechanism. Finally, I evaluate the new theory against a proposed criterion. Thus, the two main methodological approaches of this paper consist in conceptual analysis and theory development.

The third paper considers how an empirical paradigm may cast light on a conceptual debate between two conceptions of free will. Again the aim of the paper constrains the relevant methodology. I proceed from conceptual analysis to diagnose what is at stake in the debate. Other methodologies applied in this paper are conceptual development, argument and accommodation. These can be broken down into two efforts, one in the philosophical domain and one in the empirical. In the philosophical domain, I identify a process (tracking one's desires) that is a necessary premise to arguing that one conception of free will is compatible with the data. I then argue on conceptual grounds for the connection between two concepts (wholeheartedness and personal identity), and that this connection can account for the process. In the empirical domain, I explicate the connection and argue its relevance in assisting interpretation of the findings yielded by the empirical paradigm, as well as its empirical testability.

The fourth paper is concerned with arguing for an alternative interpretation of a piece of empirical data proposed in favor of a philosophical hypothesis. The necessary first methodological step in doing this is to analyze the existing interpretation of the empirical data, and to explicate the argument in which it figures. Again, the method applied can be characterized as argumentative analysis and reconstruction.

The second step is theory development pertaining to the alternative explanation of the empirical data. In order to avoid objections on conceptual grounds, the theory was developed using the conceptual framework (the higher-order thought theories) in which the original interpretation was cast. Additionally, the theory development required the alternative interpretation – in order to be a viable alternative – to be at least equal to the original interpretation in terms of explanatory power. By explanatory power, I here mean that the alternative interpretation should be able to explain at least the same amount of empirical data as the original interpretation. Lastly, in order to defend the alternative interpretation against three possible objections, I develop conceptual, as well as empirical, arguments.

The fifth paper contains three methodological components: theory development, contrastive analysis, and argument construction (conceptual as well as empirical). The theory development component consisted in developing an alternative interpretation of the empirical data. Importantly, in the development of this alternative interpretation, it was preferable that it was consistent with the general philosophical theory (i.e. the higher-order theory of consciousness) but differed only with respect to the conclusion pertaining to the specific aspect (higher-order misrepresentation) under consideration. This similarity is preferable in order to avoid begging the question against the original interpretation.

The second component is contrastive analysis, which involves comparing the original interpretation with the alternative one. This includes considering objections to the alternative interpretation, and developing replies to these in order to show that it is a viable alternative. Moreover, the contrastive analysis aimed to show the ways in which the alternative interpretation was preferable to the original one, i.e. by showing that the alternative interpretation could handle problems left unaddressed by the original interpretation. This included the development of both conceptual and empirical arguments.

The third component consisted in expanding upon the empirical plausibility of the alternative interpretation. Thus, the third component involved going over empirical findings of relevance to the alternative interpretation to validate whether the interpretation was consistent with a broader set of empirical data.

# Chapter 4 - Summary of papers

## 4.1 Paper I

The first paper, "Why the Rare Charles Bonnet Cases are not Evidence of Misrepresentation" (Kirkeby-Hinrup, 2014a), takes up the rare cases of Charles Bonnet syndrome that have been suggested as evidence for misrepresentation by the proponents of higher-order thought theory of consciousness (HOTTC). Charles Bonnet syndrome causes an individual to have complex and crisp visual hallucinations and occurs in the absence of any other cognitive disorders. Charles Bonnet syndrome can arise from a variety of causes. Predominantly it arises from ocular damage, but occasionally it is caused by damage to one or more cortical areas. In the rare cases invoked in the misrepresentation debate the cause is attributed to damage in the primary visual cortex (V1). In the argument I treat in the paper, the rare cases of Charles Bonnet syndrome are combined with empirical research into conscious visual awareness, in order to provide an argument for the occurrence of misrepresentation. The empirical research deployed in the argument suggests that the primary visual cortex plays a necessary role in the occurrence of a first-order visual state, and that conscious visual awareness requires feedback to the primary visual cortex. Thus, the rare cases of Charles Bonnet syndrome and empirical research into conscious visual awareness form the premises in an argument for misrepresentation by the proponents of HOTTC. Briefly, the argument states that if the primary visual cortex is necessary for the generation of visual first-order states and the individuals in the rare cases of Charles Bonnet syndrome lack a functioning primary visual cortex, and yet have conscious visual awareness (of the hallucinations), then presumably the conscious visual awareness they enjoy lacks relevant first-order states. To evaluate whether the rare Charles Bonnet cases are evidence of misrepresentation, I distill the argument underlying this claim. I clearly separate the premises and conclusion of the argument. This shows that the argument needs a further premise to work. However, once this hidden premise is exposed and introduced the argument is no longer sound. This is because the hidden premise entails that one of the other premises of the argument is false. Thus, the paper concludes that the rare cases of Charles Bonnet syndrome cannot be taken as evidence of misrepresentation.

## 4.2 Paper II

The second paper "How to get Free Will from Positive Reinforcement" (Kirkeby-Hinrup, 2014b), I start by noting a flaw in the concept of wholeheartedness suggested by Harry Frankfurt as the foundation for free will. The flaw is that the kind of free will we get from a wholeheartedness account is not aligned with our intuitions about what free will is. The reason it is not aligned with our intuitions about free will is that, on an account positing wholeheartedness, free will is not something possessed at all times. Thus, from the assumption that – *ceteris paribus* – a theory in line with our intuitions is preferable to one that is not, there is reason to replace the flawed concept in Frankfurt's theory. In place of the concept of wholeheartedness, I introduce a heuristic account for deliberation and decision. I then show how introspective activity can improve on these heuristics in three different ways. The first two ways are purely instrumental. The first is identification and evaluation of arguments and desires. The second is simple behavioral corrections derived from cognitively branding particular behaviors as desirable or not. While no doubt useful, these two benefits of introspective activity are unable to support a desirable notion of free will because they are inherently context dependent. The third way to apply introspection fares better in this respect. It fares better because it can influence the future deliberative behavior of an individual and is therefore not context dependent but generally applicable. The driving force in this is what I call introspective revelations. Introspective revelations provide the foundation for the individual's ability continually to optimize her cognitive behavior in relation to deliberation. What is sacrificed by rejecting the notion of wholeheartedness is that the individual can never have absolute certainty about what she wants. This certainty was obtainable in Frankfurt's account owing to the binary characteristics of wholeheartedness. In its place I introduce a self-perpetuating mechanism driven by introspective activity that provides the individual with the next best thing: increasingly good access to what she wants. The conclusion is that the propensity for introspection is the foundation for free will. This means that, in any given situation, the hallmark of free will is the propensity of the individual to determine what she wants, rather than the actual deliberative practices and decisions. A self-perpetuating mechanism driving and developing a propensity for introspection implies the commitment that this propensity can increase in strength. However, since the propensity for introspection was proposed to be the foundation of "free will" this is by extension a commitment to the idea that free will comes in degrees.

## 4.3 Paper III

The third paper, "How Choice Blindness Vindicates Wholeheartedness" (Kirkeby-Hinrup, 2015), reconsiders the Frankfurtian concept of wholeheartedness in light of the choice blindness effect from cognitive science. In the second paper, I argued that wholeheartedness should be abandoned as the focal point of a theory of free will because it was not in line with certain key intuitions about the concept of free will. In this paper, I attempt to show that the philosophical notion of wholeheartedness might serve as a useful meta-theoretical concept within the experimental paradigm of choice blindness. This in turn partly vindicates the notion of wholeheartedness by showing how it meshes with empirical sciences. I suggest that, while wholeheartedness may not be suitable as the foundation of free will, it may be fruitfully applied in another domain and need not be abandoned altogether. In the choice blindness paradigm subjects are presented with two or more alternatives (e.g. pictures of people or flavors of jam) and asked which alternative they prefer. After the subject has indicated her preference, she is presented again with the alternative she chose and asked to provide reasons for her preference of this alternative. However, in the experimental manipulation the alternative the subject is presented with after her choice is not the alternative she actually chose. The choice blindness effect is that the subjects rarely detect this bait and switch, i.e. they are "blind" to the outcome of their own choice. Most people, including prominent choice blindness pioneers, agree that there is almost certainly a limit to the sort of choices that can be manipulated in this way without the subjects detecting the ruse. I suggest that the kinds of choices that are immune to the choice blindness manipulation are those based on convictions with which the individual wholeheartedly identifies herself. In Frankfurt's account, wholehearted identification with a choice is intimately tied to the occurrence of a higher-order volition. A higher-order volition is a desire we wish to be our effective desire, a desire we wish to carry us all the way to action. From this connection it seems reasonable that exactly this kind of volition would warrant some sort of tracking of the eventual outcome. I hypothesize that it is exactly this tracking that makes wholeheartedly made choices immune to manipulation. In addition, this means that wholeheartedness has a role to play in how we define personal identity. The convictions with which I wholeheartedly identify become partly constitutive of who I am as a person and how I perceive myself. If this is right then the concept of wholeheartedness can be deployed as a meta-theoretical concept to delineate the range of the choice blindness effect.

## 4.4 Paper IV

The fourth paper, "Change Blindness and Misrepresentation" (Kirkeby-Hinrup, 2016), treats change blindness phenomenon as proposed evidence of misrepresentation by the proponents of the actualist higher-order thought theory. The change blindness phenomenon denotes the failure of experimental subjects to detect salient changes to visual stimuli. During saccades, input to the visual system is briefly cut off, leaving the subject effectively blind for the duration of the saccade. In the saccade-induced change blindness paradigm the experimenter deploys eye-tracking and specialized software to detect the onset of a saccade and cue the visual stimulus to change during the saccade. Because visual input resumes after the saccade, it is hypothesized that the post-change stimulus is present at the first-order level and the pre-saccade visual state has been overwritten in the early visual system. Cases where the subject does not report experiencing any change are explained by the subject misrepresenting her first-order visual states. In the paper, I consider an alternative interpretation of the experimental data. The alternative interpretation takes its starting point in doubting that the pre-change first-order state has disappeared completely. This interpretation I support with neuroscientific research suggesting that the pre-change first-order state may linger in cortical areas outside the early visual system. In addition to establishing that it is empirically possible that the pre-change state may survive the saccade, I argue that the proponents of the higher-order thought (HOT) theory must also accept that the pre-change state exists. They must accept that the pre-change state exists because otherwise the change blindness effect can be fully explained by the subjects being unable to compare the pre- and post-change states. I consider three possible objections to the alternative interpretation aimed at saving the change-blindness phenomenon as evidence of misrepresentation. The first objection proposes that comparison of the post-change state to the pre-change state is unnecessary for the subject to succeed in the change-blindness task, and therefore we can allow that the pre-change state does not exist, and misrepresentation can still follow. I debunk this objection by showing that mere change detection is insufficient to succeed in the change-blindness paradigms. The second objection attempts to show that even if the first-order state is still present in the subject, being conscious of it can still be counted as misrepresentation. There are two possible avenues to pursue with respect to this objection. I conclude that both of these avenues are open, but at the cost of changing the notion of misrepresentation that change blindness was supposed to be evidence for. The third objection argues that the HOT is supposed to be roughly simultaneous with the first-order state that it is about and therefore cannot be about the pre-change state. I provide two counter-

arguments to this objection. The first counter-argument points to an apparent tension between the simultaneity criterion and the possibility of misrepresentation. The second counter-argument consists in pointing to experimental evidence pertaining to the timing of processes in the visual system and arguing that the scope of what can count as simultaneous in this domain allows that the HOT can be about the pre-change state. The upshot of the article is that my alternative interpretation remains viable, and pending further evidence change blindness cannot be counted as evidence of misrepresentation.

## 4.5 Paper V

The fifth paper, "Change Blindness in Higher-Order Thought: Misrepresentation or Good Enough?" (Brinck & Kirkeby-Hinrup, *in press*), continues the examination of change blindness as evidence of misrepresentation. My co-author and I align ourselves with the alternative interpretation of the change-blindness data presented in the fourth paper. We begin by highlighting the use of empirical data in philosophy of mind, specifically in relation to David Rosenthal's actualist higher-order thought theory of consciousness. We contrapose the interpretation of the change-blindness data proposed by defenders of actualism with our alternative interpretation. We propose that the subject is in the same conscious state after the change in visual stimulus has occurred as before. This means acknowledging that the subjects are representing the pre-change visual state and, consequently, that their representations are correct.

We argue that the alternative interpretation is viable from a theoretical perspective on two grounds. First, because it deploys the conceptual framework of the actualist higher-order thought theory, it is compatible with it and cannot be rejected on principled conceptual grounds. Second, the fact that subjects can succeed in the change-blindness task seems to show that the pre-change state must be present in some form; otherwise, the subjects could not perform the comparison to the post-change state and detect what changes. We then consider two objections to the alternative interpretation. The first objection is that the post-change state cannot be represented by a higher-order thought because it has supposedly disappeared. We argue against this objection by putting forward a range of empirical findings that strongly indicate that the pre-change state may be maintained in the visual system after the stimulus switch. In summary: during an on-going activity, visual stimuli within a scene and across intervening scenes are preserved on-line in iconic memory.

The second objection is that, because actualism requires a higher-order state to be simultaneous with the lower-order state it is about, a higher-order state cannot represent the pre-change stimulus, because this is in the past. We argue against this objection in two different ways. The first is to point out that there is an apparent tension between the notion of simultaneity and the possibility of misrepresentation. We then illustrate how the simultaneity requirement can be read in two different ways, and proceed to show that both pose serious problems for actualism. The second way we argue against this objection is by defining the notion of simultaneity in terms of overlapping time segments and then showing that empirical data on the timing of conscious visual sensations can satisfy the simultaneity requirement. After treating the objections, we suggest that the alternative interpretation fits into a view of the visual system geared toward satisficing rather than truth tracking. This means that, in the absence of change detection, the visual system has no need to update the higher-order thought.

We conclude that the alternative interpretation of the change-blindness data is at least as plausible as the one advanced by actualism. By this we mean that the alternative interpretation has similar explanatory, predictive and descriptive powers. Additionally, the alternative interpretation is applicable to a wider range of change-blindness paradigms. Our investigation demonstrates the importance of examining alternative hypotheses and scrutinizing the details of empirical claims.

# Chapter 5 - Concluding remarks

## 5.1 Conclusions

In this section, I collect the conclusions of the papers. I will do this in two steps. The first step, in section 5.1.1, will consist narrowly in the conclusions that can be drawn from the substantial content of the five papers in this dissertation. The second step, in section 5.1.2, more broadly presents an overall conclusion based on the work in this dissertation.

### 5.1.1 Thematic conclusions

The papers in this dissertation yield conclusions in two distinct domains. One is the debate on the possibility of misrepresentation as posited by the proponents of the higher-order thought theory of consciousness. The other domain is the problem of free will.

The first domain pertaining to the misrepresentation debate is addressed by papers I, IV and V. In these papers, I investigated the issue of misrepresentation and two attempts to support the notion of misrepresentation by reference to empirical evidence.

From my investigations, it appears that whether the proposed empirical data constitute evidence of misrepresentation is conditional on undetermined empirical issues. Further research is needed to establish whether it is evidence of misrepresentation. One conclusion that can be drawn from this is that the proponents of the possibility of misrepresentation are faced with a challenge if they wish to maintain that the empirical data addressed in this dissertation support misrepresentation. The challenge is, on the one hand, demonstrating a way to make the argument based on the rare Charles Bonnet cases sound, and, on the other hand, showing how their interpretation of the change blindness data is preferable to the one proposed in papers IV and V.

The misrepresentation debate is interesting and, in the view of the current author, important, because it cuts to the core of a central difference between views of the

nature of our consciousness: whether conscious experience requires the instantiation of a mental state token, that is experienced, or not. Therefore, any progress we can make in this debate may help us better understand consciousness in general.

On the basis of the findings in this dissertation, one might speculate that the overall project of arguing for the possibility of misrepresentation on empirical grounds appears to be threatened. Perhaps, the possibility of misrepresentation is simply not amenable to empirical investigation. One reason to think this could be inferring inductively from the papers in this dissertation to the conclusion that examination of other proposed empirical data, or empirically based arguments, would expose similar problems, further leaving the possibility of misrepresentation underdetermined. If this is correct, then the papers here are indicative of a more severe problem for the misrepresentation debate. The reason is that the theoretical debate on the possibility of misrepresentation has largely stalled. Proponents and opponents in the theoretical debate are at an impasse. Each side of the theoretical debate advance theories that are coherent, but depend on conceptual commitments, that their opponents find implausible, counter-intuitive or absurd. If the possibility of misrepresentation is not amenable to empirical investigation, then invoking empirical data appears to bring little promise, with respect to moving forward from the theoretical impasse. I think such speculation is ill advised on several grounds. First, on the assumption that the higher-order model is correct, it would, in principle, be possible to determine whether misrepresentation occurs, if we could identify the representational relationships that obtain in a given case. Insofar as one believes that the mind can be naturalized, the process of investigating the tokening of mental states and their relations appears, in principle, to be amenable to empirical methods, albeit presumably not methods currently available. Second, there is sense in which the advancement of empirically based arguments, even if debunked, is moving the debate forward. When viewed from this perspective, a possible conclusion from this dissertation is that the introduction of empirically based arguments in fact *is* moving the debate forward. Third, while I submit that the rare cases of Charles Bonnet syndrome and the change blindness phenomenon do not hold up to scrutiny as evidence of misrepresentation, it seems premature, on the basis of what is currently known about the brain (and the small sample size considered in this dissertation), to inductively rule out that other empirical evidence will not. Thus, from the view of the current author, the empirical case for the possibility of misrepresentation is still open. However, there seems to be an important caveat with respect to the empirically based debate on the possibility of misrepresentation. The caveat is that it is unclear what would constitute empirical evidence that misrepresentation is impossible, i.e. the possibility of falsifying the misrepresentation hypothesis. Given that misrepresentation is not the normal case, it

appears, to my mind that any studies that do not find evidence of misrepresentation occurring, can likely be explained away if one so wishes. The upshot of this caveat is that the burden of proof in the misrepresentation debate appears to be on the proponents of the possibility. Importantly, I think, while it is on the proponents of misrepresentation to put forward empirical cases, where misrepresentation is hypothesized to occur, it is on their opponents to submit such empirical cases to scrutiny in order to move the debate forward.

The second domain is the problem of free will. This domain is investigated in papers II and III. The two papers explore ways in which, on the one side conceptual and on the other side empirical, arguments provide support for a theory of free will, where paper II concerns the former and paper III concerns the latter.

At the outset, it is important to address the apparent conflict between the views advanced in papers II and III. In paper II, I argue that we should reject Frankfurt's notion of wholeheartedness as the foundation of free will. In paper III, I argue that Frankfurt's notion of wholeheartedness can accommodate experimental data from the choice blindness phenomenon, and that this shows the usefulness of the notion and supports Frankfurt's theory of free will. In brief, the apparent conflict pertains to the question of whether we should reject wholeheartedness or not. However, as I will show, the conflict is merely a matter of appearance. The appearance of conflict derives from two distinct conceptions of the notion of wholeheartedness. The first conception stems from Frankfurt's theory and consists in idea that wholeheartedness is the foundation of free will. According to Frankfurt, we have free will on those occasions where we successfully execute an action with which we wholeheartedly identify. Thus, the first conception of wholeheartedness is as the *foundation of free will*. The second conception of wholeheartedness is based on its connection with personal identity that I, in the third paper, hypothesize underpins *the mechanism that tracks the outcomes of one's desires*. The reason I do not distinguish explicitly between the two conceptions in the third paper, is that, on Frankfurt's account, the two coincide. By this I mean that on Frankfurt's account, as developed in the third paper, the first conception of wholeheartedness doubles as the second conception, *to wit* the first conception plays two distinct roles. This is why, on Frankfurt's account, free will is not threatened by the choice blindness phenomenon. Crucially, the conception I reject in the second paper is the first one, i.e. the conception of wholeheartedness as the foundation of free will. The second paper does not address the second conception, and, importantly, it is the second conception that drives the conclusions of the third paper. This, of course, leaves open the question of whether the choice blindness phenomenon is a threat to the kind of free will espoused in the second paper. I will only address this question briefly here, because it deserves more comprehensive

treatment than the current context allows for. In the second paper, I propose that it is a propensity for introspection that is the foundation for free will. This propensity replaces the notion of wholeheartedness on Frankfurt's account. However, there is no commitment to this propensity manifesting itself in every situation. Furthermore, there is no commitment to correctly discovering desires, I wholeheartedly identify with, in cases where the propensity is manifested. This means we can allow for occasions, where the subject fails to track the outcomes of desires, while still maintaining that the subject has free will. This is because, on the proposed theory, free will is possessed independently of context and application. Additionally, it seems possible that the theory proposed in the second paper can incorporate the second conception of wholeheartedness, and accept that a connection between wholeheartedness and personal identity may delineate the cases where subjects detect, or fail to detect, the ruse in the choice blindness paradigm.

Now, while keeping in mind the principled epistemological separation of the conceptual from the empirical, a tentative conclusion from the two papers is that some theories of free will, e.g. Frankfurt's, can effortlessly accommodate empirical data that is perceived to threaten free will. Some theories, e.g. the theory proposed in the second paper, may accommodate the same empirical data, by working out extensions to explain the data. Finally, some theories, e.g. those that rely explicitly on deliberation and decision, are challenged by the data. It seems that the ease, with which a given theory can account for the choice blindness phenomenon, depends on what the theory identifies as the foundation of free will.

The take home message, in the domain of free will, is that there are plenty of avenues, both conceptual and empirical, yet to be investigated. Paper II in this dissertation explores the former. Paper III explores the latter.

### 5.1.2 General conclusion

This dissertation has shown the relevance and importance of assessing empirical data proposed by philosophers in support of philosophical hypotheses. In particular, I have treated the possibility of misrepresentation within the higher-order thought theories of consciousness and the choice blindness phenomenon in relation to the problem of free will. As for the possibility of misrepresentation, I have shown that it is highly doubtful that two pieces of empirical data unequivocally show what they have been purported to. One possible upshot of this is that the suggested practice of double checking the abductive arguments based in empirical data proposed in favor of a philosophical hypothesis shows promise and is worthy of further pursuit. From the

view of the current author this practice is invaluable to the future development both in philosophical research and the involved empirical sciences. From this point of view, allocating attention to proposed empirical data in favor of a philosophical theory is worthwhile, irrespective of whether the theory in question is one's own. However, it seems natural that when it comes to the practice of advancing empirically based arguments against a philosophical hypothesis, the burden of labor will fall on the opponents of a given theory. Reasonably and charitably, one would expect researchers to propose arguments in favor of a hypothesis only if they believe those arguments to be valid, and believe the hypothesis is worth arguing for.

Regarding free will, I have shown how empirical evidence may suggest that intuitions about the nature of free will are mistaken, and that this may help to evaluate arguments formed on the basis of these intuitions. Additionally, this dissertation has contributed to conceptual work within the area of free will by developing a new theory in this area.

## 5.2 Viewing the papers from the meta-perspective

The meta-perspective considers different ways in which one might approach work within the empirical turn, as well as serving to underscore the value of this practice as perceived by the current author.

In this section, I will provide a view of the papers in this dissertation in light of the discussions of the empirical turn in the previous chapters. In section 5.2.1, I will consider how the papers in this dissertation, viewed collectively, relate to the meta-perspective. In the following section (5.2.2), I will consider the papers individually in light of the meta-perspective. However, before continuing, it is important to underscore that this marriage between the papers and the meta-perspective is primarily *post hoc*. While the papers on misrepresentation, in particular, touch on many of the issues of the meta-perspective discussed in the previous chapters, the papers were not written with this perspective explicitly in mind. Instead, the meta-perspective grew out of the papers, as it were, through the work on them and by retrospectively considering their relation to other work in the field.

### 5.2.1 The papers viewed collectively

The papers in this dissertation show at least two things when viewed from the meta-perspective.

The first is that the practice of re-assessing empirical evidence proposed in favor of philosophical hypotheses is viable, fruitful and important. This is shown because the three papers on the misrepresentation debate demonstrated that the proposed evidence did not succeed in showing what it was purported to. This, in turn, points to the central caveats with respect to making inference to the best explanation arguments in favor of philosophical hypotheses on the basis of empirical evidence. At the same time, the papers on misrepresentation, as well as the third paper pertaining to choice blindness, can be seen as showing how empirical evidence may move philosophical debates forward.

The second is that, while philosophers are, understandably, mainly concerned with philosophical endeavors and therefore how empirical research may influence these, the interdisciplinary interaction need not be unilateral. Considering empirical evidence in the papers led to concepts relevant for empirical researchers, and empirical predictions and theory development that may interest empirical researchers, as well as philosophers. This indicates that the empirical turn is of mutual benefit for philosophers and researchers in empirical sciences alike.

### 5.2.2 The papers viewed individually

Because it is concerned with reassessing proposed empirical data, when considered from the meta-perspective, the first paper (Kirkeby-Hinrup, 2014a) may be considered as belonging to the third practice discussed in section 2.2.3. In the paper, I am not disputing the understanding of the empirical data. Rather, I object to the argument in which the data figure as a premise. In the paper, I argue that there is an ambiguity in the interpretations of two pieces of empirical data. This appears to be an example of conceptual mapping gone awry, where concepts that appear to denote the same phenomenon on the surface are shown, upon scrutiny, to diverge in their extensions. In the paper, I explicitly address Ned Block's suggestion that theories of (phenomenal) consciousness should be merited by the extent to which they mesh with neuroscientific data. This, indirectly, touches on the discussion about inference to the best explanation in the meta-perspective.

The second paper (Kirkeby-Hinrup, 2014b) can be seen as an example of philosophical work taking inspiration from empirical sciences. In the paper, I present a novel philosophical theory on the basis of a well-established and uncontroversial empirical phenomenon. Importantly, the empirical phenomenon I deploy in the theory serves to fill a role previously played by a purely philosophical concept that was deemed inadequate for conceptual reasons. Because the empirical phenomenon is

broadly recognized and well defined among both philosophers and empirical scientists, it appears to avoid the worries pertaining to conceptual mapping mentioned in section 2.2.1.3.

The third paper (Kirkeby-Hinrup, 2015) is an example of conceptual clarification as discussed in section 2.3.1. This is because I propose that a philosophical concept may be of use within a specific group of empirical paradigms to describe certain limitations of those paradigms. Additionally, this paper may be seen as an example of explicit input, discussed in section 2.3.3, on two points. The first point is that it offers up a philosophical concept that may be of use to empirical scientists. The second point is that the paper suggests a way to investigate empirically my proposed connection between two philosophical concepts.

The fourth paper (Kirkeby-Hinrup, 2016) might be categorized as an example of "philosophical replication" (the third practice discussed in section 2.2.3), whereby one re-assesses the empirical data and the arguments proposed in favor of a philosophical hypothesis. I present an alternative interpretation of the empirical data, and argue that the alternative interpretation is more plausible on both theoretical and empirical grounds. This means that, in addition to reassessing and objecting to a piece of empirical data, this paper could be seen to propose an inference to the best explanation (section 2.2.1.1) in favor of the alternative interpretation of the empirical data. In the paper, I explicitly address the role empirical data plays in inference to the best explanation, in the debate between the higher-order thought theory and its opponents. Additionally, I mention the caveat that it is mistaken to believe that empirical data can have any bearing on conceptual issues.

As for the fifth paper (Brinck & Kirkeby-Hinrup, *In press*), when viewed from the meta-perspective, the philosophical and empirical arguments provided in favor of our alternative account can be seen as akin to an inference to the best explanation argument (section 2.2.1.1). Because the fifth paper is significantly more empirically minded than the fourth paper, there is a sense in which the relation between the two may be seen as an example of conceptual alignment (section 2.2.1.3) between philosophical theory and empirical data. In this sense, the third article provides support for the second article by showing that the theoretical concepts and processes invoked in the second article have empirical counterparts, and thereby that the alternative interpretation is empirically plausible. The fifth paper explicitly addresses some of the topics included in the meta-perspective. First, we address how empirical evidence is deployed to inform philosophy of mind. Second, we explicitly submit that assessing proposed empirical evidence from an independent perspective has a critical role to play in empirically minded philosophy. Third, we highlight the principled

distinction between the empirical and the conceptual. Fourth, we address caveats related to inference to the best explanation on the basis of empirical evidence.

## 5.3 Future research based on the papers

On the basis of papers I, IV and V, there is interesting work to be carried out re-assessing additional empirical evidence proposed in favor of the possibility of misrepresentation. More generally, the papers may be viewed as supporting the practice of re-assessing empirical evidence proposed in favor of hypotheses in philosophy of mind as a promising venue of future work.

Concerning the fourth and fifth papers, future areas of research could consist in the continued substantiation of the alternative interpretation of the change-blindness phenomenon. This might consist in reviewing further empirical findings of possible relevance to the alternative interpretation, as well as considering its philosophical implications and applications outside the domain of higher-order misrepresentation.

With respect to the second paper and the proposed theory of free will, there may be future work related to addressing incoming objections to the theory. The character of such objections, and the work they may prompt, is open to speculation. Additionally, there are different directions future investigations may take depending on which aspects of the theory one finds interesting. Given that the theory is neutral on the metaphysical and normative issues, there is room for philosophically interesting extensions to the theory to be worked out in these domains. One avenue of future research I find intriguing, in relation to the proposed theory, is an effort in conceptual mapping between the mechanisms driving introspection (and the self-perpetuating propensity for introspection) and empirical data from cognitive science and neuroscience. Another interesting area to expand upon is the so-called "selection problem". In my theory the selection problem pertains to the proposed mechanism that selects which mental states become conscious (in the paper, I use Dennett's term 'consideration generator' to denote this mechanism). The description of this mechanism is left undescribed in the proposed theory because what is relevant for the theory is mainly the set of mental states that are available to it (i.e. the set from which it selects). Thus, hypotheses of both philosophical and empirical kinds on how this mechanism accomplishes its task would be relevant additions and good candidates for future efforts in the development of this theory.

As for the third paper, there appears to be a least two possible avenues to take when it comes to future research. The first avenue is conceptual. It has been suggested

(Gåvertsson, 2016) that my argument, that theories of free will relying on wholeheartedness are not threatened by the choice blindness phenomenon, can be applied to other accounts of free will. Investigating such wider applicability would be interesting. The other avenue one might take is empirical. This consists in empirically testing my proposed connection between wholeheartedness and personal identity by modifying the choice blindness paradigm along the lines I suggest in the paper.

## 5.4 Future research within the empirical turn

In this section, I will briefly consider four commonsense tenets directed mainly at philosophers working within the empirical turn. As with the meta-perspective, these tenets are derived *post hoc* from the papers, and have grown out of my considerations of the many caveats discussed throughout the meta-perspective.

I call these tenets commonsense because I do not expect them to be contested. If someone were to contest them, it would probably be on accusation of triviality. Still, I submit, the commonsense tenets are useful to keep in mind for two reasons. First, they serve as guidelines presumed to benefit the way research is conducted in relation to the empirical turn. Second, the tenets are useful to analyze and evaluate work where empirical data intersects with philosophy. Before I proceed to describing them, I want to mention one additional tenet that could have been included here. This tenet prescribes staying informed of empirical data relevant to one's research. However, the empirical turn is predicated upon this being the standard practice. Thus, this tenet is actually a pre-condition for the empirical turn and including it would be superfluous.

*Tenet 1: Propose empirical data in support of your theory*

Insofar as one is engaged in the empirical turn, the attempt to create a mesh between one's philosophical theory and empirical data is an important task. Importantly, this tenet involves more than merely staying informed of empirical research pertaining to your area of study, because it urges also applying this information. One might think that this task is only feasible if the kind of theory one espouses can invoke concrete and specific data, as is the case with, for example, the naturalist theories of consciousness. Supposedly, the task becomes more difficult in the domain of free will, where most empirical data of relevance take the form of general gestures toward either determinism or indeterminism, or conscious processes being causally efficacious as opposed to epiphenomenal. However, as evidenced by the second and third papers, it is possible to participate in the free will debate on the basis of concrete empirical

phenomena. At any rate, the idea is that, to make the most of the empirical turn, it is advisable actively to seek out empirical data of relevance to one's views.

*Tenet 2: Assess empirical data of relevance to competing theories*

Given the philosophical tradition of constructing complex theoretical counterexamples, finding logical flaws and debating the concepts invoked by competitors, one would expect this practice to be carried over to proposed empirical support. Surprisingly, philosophers have not significantly extended their work on counter-arguments to include empirical ones. Perhaps one explanation for this is that the empirical turn is still in its infant years, and that philosophers interested in the mesh with empirical sciences have devoted most of their efforts to the first tenet. Regardless, the scientific burden of scrutinizing the work of other researchers cannot be left exclusively on the shoulders of editors and reviewers of scientific journals. Assessing empirical data proposed by competitors has a crucial role to play in the empirical turn. For instance, because the mesh between philosophy and empirical sciences relies essentially on interpretation and each philosopher has his own axe to grind, there is a risk of errors arising from cognitive biases. In addition, logical mistakes or reasoning errors are as likely to occur in the interpretation of empirical data as in other practices. Finally, that a theoretical prediction fits with a piece of empirical data does not automatically entail that the underlying theory is true. There may be alternative explanations of the result, perhaps even interpretations that support a competing hypothesis. In relation to the second tenet, one should remember that it not only pertains to assessing empirical data proposed in favor of competing theories, but additionally to proposing empirical data that present challenges to competing theories. Attempting to identify and bring such data to light is part of the progress of the empirical turn, and is of course especially relevant when considering competing theories.

*Tenet 3: Consider whether and how your philosophical work may inform empirical research*

Not only philosophy stands to gain from the interdisciplinary venture with the empirical sciences. Much philosophical effort has been put into developing and communicating conceptual analyses of concepts central to related empirical sciences. As a result, neuroscientists these days have a much more detailed and precise understanding of concepts such as consciousness, rationality, phenomenality, introspection, etc. than just a few decades ago. Philosophers are mainly to be credited for this. To provide input actively to empirical sciences perpetuates the empirical turn, and brings the relevance of philosophy in general, and your input specifically, into the minds of empirical scientists.

*Tenet 4: Take non-contentious concepts as a starting point*

Before it is feasible to draw inferences from empirical data one needs, as a minimum, an interpretation that maps the relevant concepts of the empirical data and the philosophical theory to each other. Obviously, the preliminary conceptual mapping influences the kinds of inference available to be made from the empirical data. Furthermore, the philosophical predilections of the interpreter are bound to influence the mapping of concepts between theory and empirical data. Such mapping is a necessary and important aspect of the empirical turn. However, care needs to be taken when doing so. It seems that the conceptual commitments one has in the philosophical domain are likely to influence the way in which one interprets the empirical data. Importantly, this is not a question of bias in the interpretation. Rather, it is a natural consequence of the conceptual and theoretical commitments of the interpreter. It would be unfair to expect researchers, when interpreting relevant empirical data, not to make use of the concepts they think best describe and categorize the phenomenon under investigation. Often, since this mapping influences the inferences one can make from the empirical data, disagreement on the mapping may show up in the inferences, i.e. the conceptual disagreements can migrate into the interpretations of the empirical data. This does not pose a huge threat to the empirical turn as such; we can still draw inferences to the best explanation on the basis of how well competing theories mesh with empirical data. Furthermore, the conceptual framework of a theory is often a part of the supposedly best explanation. It is only where we hope that appealing to empirical data can arbitrate and move forward debates on specific issues that have reached an impasse on conceptual and theoretical grounds that this may become a problem. In these cases we risk encountering a new impasse over the correct interpretation of the empirical data. Pre-empting this is the motivation for the fourth tenet. One should aim, when employing empirical data, to stick with the consensus interpretation (if one exists). The fewer concepts esoteric to the philosophical theory one attempts to mesh with empirical data deployed in the interpretation the better.

# References


Bayne, Tim. (2008). The unity of consciousness and the split-brain syndrome. *Journal of Philosophy, 105*(6), 277-300.

Bayne, Tim, & Chalmers, David. (2003). What is the unity of consciousness. In: Cleeremans, Axel (Ed.), *The unity of consciousness: Binding, integration, dissociation,* 23-58. Oxford University Press.

Björnsson, Gunnar. (2016). Outsourcing the deep self: Deep self discordance does not explain away intuitions in manipulation arguments. *Philosophical Psychology, 29*(5), 637-653.

Block, Ned. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227-247.

Block, Ned. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30*(5-6), 481-499.

Block, Ned. (2008). Consciousness and cognitive Access. *Proceedings of the Aristotelian Society, 108*(1pt3), 289-317.

Block, Ned. (2014). Rich conscious perception outside focal attention. *Trends in Cognitive Sciences, 18*(9), 445-447.

Bonn, Gregory B. (2013). Re-conceptualizing free will for the 21st century: acting independently with a limited role for consciousness. *Frontiers in Psychology, 4*.

Brinck, Ingar. (2001). Attention and the evolution of intentional communication. *Pragmatics & Cognition, 9*(2), 259-277.

Brinck, Ingar. (2004). The pragmatics of imperative and declarative pointing. *Cognitive Science Quarterly, 3*(4), 1-18.

Brinck, Ingar. (2015). Understanding social norms and constitutive rules: Perspectives from developmental psychology and philosophy. *Phenomenology and the Cognitive Sciences, 14*(4), 699-718.

Brinck, Ingar, & Kirkeby-Hinrup, Asger. (*In Press*). Change blindness in higher-order thought: Misrepresentation or good enough? *Journal of Consciousness Studies*.

Brogaard, Berit. (2011). Are there unconscious perceptual processes? *Consciousness and Cognition, 20*(2), 449-463.

Chalmers, David. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies 2*(3), 200-219.

D'Aloisio-Montilla, Nicholas. (*forthcoming*). Imagery and overflow: We see more than we report. *Philosophical Psychology*.

Deutsch, Max. (2010). Intuitions, counter-examples, and experimental philosophy. *Review of Philosophy and Psychology, 1*(3), 447-460.

Di Pellegrino, Giuseppe, Fadiga, Luciano, Fogassi, Leonardo, Gallese, Vittorio, & Rizzolatti, Giacomo. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research, 91*(1), 176-180.

Dowe, David L., Hernández-Orallo, José, & Das, Paramjit K. (2011). Compression and intelligence: social environments and communication In: Schmidhuber, Jürgen, Thórisson, Kristinn R., & Looks, Moshe. (Eds.) *Artificial General Intelligence*, 204-211. Springer.

Feltz, Adam. (2009). Experimental philosophy. *Analyse & Kritik, 31*(2), 201-219.

Fleming, Stephen M., Dolan, Raymond J., & Frith, Christopher D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 367*(1594), 1280-1286.

Frankena, William K. (1939). The naturalistic fallacy. *Mind, 48*(192), 464-477.

Frankfurt, Harry G. (1958). Peirce's notion of abduction. *Journal of Philosophy, 55*(14), 593-597.

Gallese, Vittorio. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1480), 659-669.

Gallese, Vittorio, & Goldman, Alvin. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493-501.

Gallese, Vittorio, & Sinigaglia, Corrado. (2011). What is so special about embodied simulation? *Trends in Cognitive Sciences, 15*(11), 512-519.

Goldberg, Ilan I., Harel, Michal, & Malach, Rafael. (2006). When the brain loses its self: prefrontal inactivation during sensorimotor processing. *Neuron, 50*(2), 329-339.

Gomes, Gilberto. (1999). Volition and the readiness potential. *Journal of Consciousness Studies, 6*(8-9), 59-76.

Gordon, Robert M. (1986). Folk psychology as simulation. *Mind & Language, 1*(2), 158-171.

Grush, Rick, Jaswal, Liberty, Knoepfler, Justin, & Brovold, Amanda. (2015). Visual adaptation to a remapped spectrum. *Open MIND.* Frankfurt am Main., MIND Group.

Gåvertsson, Frits P. W. (2016). Response to Asger Kirkeby-Hinrup. *Organon F, 23*(1), 125-127.

Harman, Gilbert H. (1965). The inference to the best explanation. *The Philosophical Review, 74*(1), 88-95.

Hohwy, Jakob. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology, 3*(96).

Jackson, Frank. (1982). Epiphenomenal qualia. *The Philosophical Quarterly, 32*(127), 127-136.

Jackson, Frank. (1986). What Mary didn't know. *Journal of Philosophy, 83*(5), 291-295.

Kirkeby-Hinrup, Asger. (2014a). Why the rare Charles Bonnet cases are not evidence of misrepresentation. *Journal of Philosophical Research, 39*, 301-308.

Kirkeby-Hinrup, Asger. (2014b). How to get free will from positive reinforcement. *SATS, 15*(1), 20-38.

Kirkeby-Hinrup, Asger. (2015). How choice blindness vindicates wholeheartedness. *Organon F, 22*(2), 199-210.

Kirkeby-Hinrup, Asger. (2016). Change blindness and misrepresentation. *Disputatio, 8*(42), 37-56.

Knobe, Joshua. (2003a). Intentional action and side effects in ordinary language. *Analysis, 63*(279), 190-194.

Knobe, Joshua. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*(2), 309-324.

Knobe, Joshua, & Prinz, Jesse. (2008). Intuitions about consciousness: Experimental studies. *Phenomenology and the Cognitive Sciences, 7*(1), 67-83.

Kouider, Sid, De Gardelle, Vincent, Sackur, Jérôme, & Dupoux, Emmanuel. (2010). How rich is consciousness? The partial awareness hypothesis. *Trends in Cognitive Sciences, 14*(7), 301-307.

Kriegel, Uriah. (*forthcoming*). Beyond the neural correlates of consciousness. In: Kriegel, Uriah (Ed.), *Oxford handbook of the philosophy of consciousness*: Oxford University Press.

Lamme, Victor A. F. (2003). Why visual attention and awarenes are different. *Trends in Cognitive Sciences, 7*(1), 12-18.

Lamme, Victor A. F. (2004). Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness. *Neural Networks, 17*(5-6), 861-872.

Lau, Hakwan. (2007). A higher order bayesian decision theory of consciousness. *Progress in Brain Research, 168*, 35-48.

Lau, Hakwan, & Brown, Richard. (*Forthcoming*). The Emperor's New Phenomenology? The Empirical Case for Conscious Experiences without First-Order Representations. In: Pautz, Adam. & Stoljar, Daniel. (Eds.), *Festschrift for Ned Block*: MIT Press.

Lau, Hakwan, & Rosenthal, David. (2011). Empirical support for higher-order theories of conscious awareness. *Trends Cognitive Sciences, 15*(8), 365-373.

Levine, Joseph. (1983). Materialism and qualia: The explanatory Gap. *Pacific Philosophical Quarterly, 64*(4), 354-361.

Levy, Benjamin J., & Anderson, Michael C. (2012). Purging of memories from conscious awareness tracked in the human brain. *The Journal of Neuroscience, 32*(47), 16785-16794.

Libet, Benjamin. (2004). *Mind time: The temporal factor in consciousness*: Harvard University Press.

Malach, Rafael. (2011). Conscious perception and the frontal lobes: comment on Lau and Rosenthal. *Trends Cognitive Sciences, 15*(11), 507; author reply 508-509.

Mayo, Michael J. (2003). Symbol grounding and its implications for artificial intelligence. *Proceedings of the Twenty-sixth Australasian Computer Science Conference on Research and Practice in Information Technology,* 55-60. Australian Computer Society Inc..

McGinn, Colin. (1991). *The problem of consciousness*. Blackwell.

Miller, George A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences, 7*(3), 141-144.

Nagel, Thomas. (1974). What is it like to be a bat. *Philosophical Review, 83*(4), 435-450.

Nichols, Shaun, & Knobe, Joshua. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous, 41*(4), 663-685.

Nilsson, Nils J. (2005). Human-level artificial intelligence? Be serious! *AI Magazine, 26*(4), 68-75.

O'Regan, Kevin j., Myin, Erik, & Noë, Alva (2005). Sensory consciousness explained (better) in terms of bodiliness and grabbiness. *Phenomenology and the Cognitive Sciences*, *4*(4), 369-387.

Overgaard, Morten. (2015). *Behavioural Methods in Consciousness Research*. Oxford University Press.