Magister Essay, Theoretical Philosophy – Stockholm University

A Choice-Blind Inner Sense

Can the “choice blindness”-experiments be used to test the inner sense-theories of self-knowledge?

Johan G:son Berg

2015, fall semester

Personal number: 19890328-0058

E-mail: johangsonberg@outlook.com

Supervisor: Åsa Wikforss

Summary

The choice blindness-experiments suggest that the inner sense-theory is incorrect, and that we do not have an inner sense with direct access to our beliefs. These experiments use deception to show that many people can within minutes, and without thinking that anything is wrong, self-ascribe two incompatible beliefs. If we had an inner sense, the argument goes, we would use this inner sense to find that same belief every time. This would mean that errors, such as self-ascribing incompatible beliefs, should be impossible or rare. Since they are not, the inner sense theory cannot be correct.

I argue that the inner sense-theory can avoid the defeating implications from the choice blindness-experiments by adding details concerning how the inner sense operates. Furthermore, it is unclear if we can draw any strong conclusions from the fact that we can be deceived, and that using deception as a method might be a problematic. I suggest some other ways of testing the inner sense-theory, and conclude that it is hard to tell what any result might mean without a clear idea of both the specifics of the theory as well as the nature of belief.

Keywords: self-knowledge, inner sense-theory, theory-theory, choice blindness

Content

1. Introduction

A standard way to find out what people think about something, be it from different kinds of jam or political parties, is to issue a survey. In these, people are instructed to tell us what they think by answering questions or ranking statements. We generally do not concern ourselves with how people go about when they fill out this kind of forms, and simply assume that their statements accurately reflect their beliefs on these matters – that they have *self-knowledge*. In this essay, I will look more closely at one model of how people are able to report on what they believe: the inner sense-theory. This theory, although there are variations to it, states that we have an inner sense that provides us with a direct access to our beliefs – we "look inwards" or "introspect". Disregarding different interpretations of "direct access" for the moment, it seems natural to assume that this means that we usually are right about what we believe, when we use the inner sense. This is, as a comparison, different from when we try to find out what other people believe, where we often can be wrong if we try to only go by non-verbal clues, such as facial expression or behaviour.

The purpose of this essay will evaluate two empirical experiments that seems to suggest that the inner sense-theory is incorrect, and that we do not have an inner sense with direct access to our beliefs. These experiments, the *choice blindness* experiments, use deception to show that within minutes, and without thinking that anything is wrong, we can self-ascribe two incompatible beliefs. They also show that this is not only possible in a few cases, but for most of the test subjects. If we had an inner sense, the argument goes, we would use this inner sense to find that same belief every time. This would mean that errors, such as self-ascribing incompatible beliefs, should be impossible or rare. Since they are not, the inner sense theory cannot be correct.

I argue that the inner sense can avoid the defeating implications from the choice blindness-experiments by adding details concerning how the inner sense operates. Furthermore, it is unclear if we can draw any strong conclusions from the fact that we can be deceived, and that using deception as a method might be a problematic. I suggest some other ways of testing the inner sense-theory,

and conclude that it is hard to tell what any result might mean without a clear idea of both the specifics of the theory as well as the nature of belief.

I will start by giving a brief background of the field of self-knowledge, and then focus on three variations of the inner sense theory of self-knowledge. I will also give an overview of the competing *theory*-theory, which provides an alternative account of how self-knowledge is acquired. This will be used as a comparison to the inner sense theory.

After that I will describe how the choice blindness-experiments are executed, what the results were and in what way they are thought to be incompatible with an inner sense theory. Finally I will discuss how an inner sense theory could attempt to avoid these potential problems, and discuss whether this is a viable strategy.

2. Self-knowledge

2.1 Background

Since the term “self-knowledge” is sometimes used to refer to several different phenomena, it will be useful to start with a definition and a delimitation. A *belief* is the propositional attitude of holding a proposition true. *Self-knowledge* in this paper will concern knowledge of one’s own beliefs; to have self-knowledge is to have a true and justified second order belief about one’s own first order belief. The theories discussed in this essay differs on how to treat the justification criterion of self-knowledge. There are other usages of the term “self-knowledge”. It can refer to knowledge of one’s own character traits, cognitive mechanisms or subconscious states, for instance in the context of psychology. Self-knowledge can also concern knowledge of one’s own sensations, or knowledge of other propositional attitudes such as doubting, fearing or hoping. However, none of these other types of self-knowledge are the subject of this essay.

Philosophers often agree that self-knowledge is special and different from other knowledge. However, they also often disagree on in what way it is special. There are *rationalists* such as Tyler Burge and Richard Moran who claim that self-knowledge of belief is special because of the important role of second-order beliefs in critical reasoning (Gertler, 2011, pp. 67-68). There are on

the other hand philosophers such as Sydney Shoemaker and Eric Schwitzgebel who claim that self-knowledge is special because of the nature of belief. Their *partial constitution* view states that a true believer self-ascribes the belief, and that having a second order belief is a part of what it is to have a first-order belief (Schwitzgebel, 2011).

Self-knowledge might also be special in an epistemic way. This would mean that it is more secure, better justified, achieved by a special method, or any of these in combination. The strongest claim concerning epistemic security is that self-knowledge is secure to the extent that it is both *infallible* and *omniscient*. The first claim simply says that we never have erroneous second order beliefs. The second claim means that all first order beliefs are always accompanied by true second order beliefs. Neither of these claims is commonly endorsed today since we might from time to time encounter examples where we realize that we had a belief we were not aware of, and where we realize that we had a different belief than the one we thought (Gertler, 2014, pp. 62-63). An example might be: I have the belief that I believe it to be hot outside, but realize I took my warm jacket when I left the house, and that it actually was cold outside – which might suggest that I (unless I simply made a mistake) really believed it to be cold outside. Those who claim that self-knowledge is epistemically special commonly claim that we do not *usually* have false second order beliefs, and that *most* beliefs are at least *available* to self-knowledge.

An example of a special method that could be used in self-knowledge is provided by the inner-sense. An inner-sense theory claims that we use introspection, “look inwards”, using a particular capacity, i.e. the inner sense. According to this theory we have *privileged access* to our own mental states – we know them in a way that no one else does. This theory is proposed by David Armstrong (1971), Stephen Nichols together with Shaun Stich (2003) and Alvin Goldman (2006). They claim that we form our second-order beliefs by using the inner sense to detect our first-order beliefs. This knowledge is thus mediated by some sort of mechanism, and hence no inner sense theory has to claim that self-knowledge is infallible. However, it is usually assumed to be more

secure than other kinds of empirical knowledge and that we are more justified in our second order beliefs acquired by the inner sense than in other beliefs about the world.

Some theories of self-knowledge theories deny that self-knowledge is special in any sense, epistemic or otherwise. An example is the *theory*-theory which claims that we acquire knowledge of our own beliefs by using the same methods we use to get knowledge of other peoples' beliefs. One strong proponent of this is Alison Gopnik, but several other philosophers, such as Daniel Dennett and Jerry Fodor (to mention two), could be understood as supporting some aspects of the theory-theory (Gopnik, 1993). She suggests that we have a *theory* connecting behaviour with mental states, such as "shouting" with "being angry", and we know our intentional mental states by observing our behaviour and applying this theory.

The inner sense theories claim, by contrast, that we have a *direct* access to our beliefs. This is not to be confused with *instant* access; the inner sense-theory need not say anything regarding the time required to use this sense. To say that the access is direct is to say that we have access to the belief itself, rather than any of the consequences that having the belief might bring about. Furthermore, direct access does not mean that the access is not *mediated* by something – it is mediated by the inner sense. However, the access is *non-inferential*, meaning that while it is mediated by the inner sense, the inner sense does not use other beliefs as evidence for the second order beliefs. The theory-theory, as a comparison, claims that we have indirect access to our beliefs, and that we use behavioural clues as evidence to infer our second order beliefs. Both the inner-sense theory and the theory-theory are in one sense or another *realists* about beliefs, and consider them to be mental states that "are there" for us to know about. This is why, for instance, the theory-theory is not a type of behaviourism (Gopnik, 1993, p. 12).

One dividing line between the inner sense-theory and the theory-theory concerns how to interpret, account for or explain the empirical evidence available from experimental psychology. Much of the debate on this point is about whether the evidence supports the purported difference between *first-person knowledge*, knowledge of your own mind, and *third-person knowledge*,

knowledge of someone else's mind. In this paper, the disputed question is whether or not we can interpret apparent mistakes in self-attribution of beliefs as arguments against a privileged access account of self-knowledge, such as the inner sense theory.

Section 2.2 provides a general outline of shared characteristics of the different versions of the inner sense theory. This outline will provide the base for the explanation of why the "choice blindness"-experiments could be incompatible with the inner sense theory. Then I describe three different versions of the inner sense theory: Armstrong's, Nichols and Stich's, and Goldman's versions. The variations between them will give slightly different explanations of the empirical results discussed. Finally, in section 2.3, I present an account of Gopnik's theory-theory, which could possibly present a better explanation of these results, since it allows for more frequent errors.

2.2 The inner sense theory

The inner sense theory has been traced back to ideas of René Descartes, John Locke and Immanuel Kant (Gertler, 2011, pp. 39-47). While it might be possible to dispute how much of their ideas is left in the modern versions of the inner sense theory, they all took note of our ability to consider the content of our own minds, and supposed that it appeared to be done in a fashion similar to vision. This idea of introspection, quite literally "looking inwards", is a linchpin in all versions of the inner sense theory. Taking this idea of "seeing" literally also entails that introspection sometimes fails, just as our senses sometimes fail us. It also suggests that we do not ordinarily make mistakes when introspecting, just as our senses seem to serve us well in most situations.

According to all versions of the inner sense theory we form second-order beliefs by "looking inwards", in one fashion or another, at our beliefs. None of the theories discussed below is very explicit in describing how a person *reports* her beliefs. For our present purposes this is slightly problematic, since we can only test belief reports. It might then be possible for a person, according to an inner sense theory, to form true second-order beliefs, but fail to express them accurately. However, even if this might happen on some occasions, the connection between our second-order beliefs and our ability to express them would arguably have to be strong, since we have no

indication of this type of failure happening very often. The following theories all seem to assume that there is not a problematic step between having a belief and reporting on it, implying that we can treat the self-ascription of belief directly as the expression of a second order belief.

The three versions describe the process acquiring self-knowledge somewhat differently, using different concepts. However, all three theories use a notion of an "inner sense", which Armstrong calls a self-scanner and Nichols and Stich call a monitoring mechanism, and I will use "inner sense" to mean any of these. Similarly, I will use "introspection" as a broad term (with the exception of section 2.2.3) referring to the act of using this inner sense to create second-order beliefs. When I sometimes say "*the* inner sense-theory" I am referring to aspects which all three theories share, in situations where it is possible to treat them as one theory.

2.2.1 Armstrong's version

In his book "A Materialist Theory of the Mind", first published in 1968, David M Armstrong presents what he thinks of as a general image of the workings of the mind. The fifteenth chapter of his book is devoted to introspection, and is the base for his inner sense theory (Armstrong, 1971, pp. 323-338).

Structurally, the chapter comes directly after his treatment of perception and bodily sensations, and he takes introspection to be closely similar to sense perception. Perception is, to Armstrong, a mental event with its intentional object outside the mind, and introspection is simply a mental event with another mental event of the same mind as its intentional object. This introspection may in turn be the intentional object for another mental event, and so on, up to the limitations given by the brain. Since a mental state cannot be aware of itself, introspective awareness requires something else to be aware of that mental state, which makes a regress necessary. However, this chain of introspection has to end up in an introspection that is not introspected in order to avoid a vicious infinite regress (Armstrong, 1971, p. 324). Introspection thus needs to include some direct awareness of the mental state. To Armstrong, the inner sense is similar to proprioception, the perception of our body, which does not involve a sense-organ in the

fullest sense, unlike vision or hearing. When we perceive our body or a mental state there is nothing we perceive *with* (Armstrong, 1971, p. 325). Armstrong's description of the inner sense, or self-scanner, appears in some ways to be more of a characteristic of the workings of the mind than a sense. Introspection is for instance thought to be direct and non-inferential, and the scanner is not supposed to be thought of as a "'search light' that makes contact", in the way that we "make contact" with objects when using our eyes (Armstrong, 1971, p. 326). It is more like when we perceive a part of our body, say the tongue or a little toe, which we normally are not aware of but can bring to mind when we want to.

Since the inner sense is not infallible, it is important to spell out in what situations it might fail. Armstrong says that we are able to be misinformed about our own mental states in a similar fashion as we may be misinformed about the objects we perceive by our other senses. Just as there are many features of our environment that we do not perceive, there is much in our mind that we are not introspectively aware of. And just as we can mistake a bush for a bear, we can mistake fear for anger when we are introspecting. In this case, the false introspection that we are angry is brought about by the mental state of fear (Armstrong, 1971, p. 330). This is equal to denying the omniscience and the infallibility claim.

There are two other cases where Armstrong opens up for the possibility that there can be a gap between when introspecting. I can, according to Armstrong, perceive myself, one of my mental states and at the same time question if I am aware of this mental state. Armstrong imagines the case where we have acquired information which we do not like, and therefore are reluctant to accept (Armstrong, 1971, p. 332). In this case our mind is divided, we have a "split consciousness", and one part of our mind questions whether another part of our mind has accepted this mental state. What Armstrong means by "accepted" is unclear, but it seems that it would have to involve the formation of a second order belief.

There can also be cases where we by conscious effort refrain from forming a second order belief. Armstrong imagines the case where a person uses the inner sense but does not form a second

order-belief about the belief that the inner sense appears to find. It is not clear exactly how we are to understand what “appear” might mean here. This might be analogous to the ability to override our perceptions when we form extensive experience and with strong justification know that we are being deceived – perhaps as with optical illusions or phantom limbs. Then an introspection-case might be as follows: when I employ the inner sense and appear to find the belief “poor people are dangerous”, I do not form a second-order belief but, instead use my experience and form the belief: “I believe that I am afraid of people that are not like me”. However, this requires that we are able to correct our introspective reports, writes Armstrong, and this is difficult to achieve: “In a future where far more is known than at present about the workings of the brain, it would be possible to be quite sure that certain introspections were illusory. I might appear to myself to be angry, but know myself to be afraid.” (Armstrong, 1971, p. 328) The choice blindness-experiments might be just the kind of experiments that could do this in the case of belief.

In the years passed since the publication of Armstrong’s book in 1971, the evidence showing that we do indeed make quite a lot of introspective errors has accumulated. Armstrong is vague on why we would make this kind of errors: if the awareness is supposed to be direct and non-inferential, then it seems we should not be able to make these kinds of errors at all, unless – perhaps – we have a serious mental disorder. How could we form the mistaken belief that we are angry when introspecting our fear? While we might use our experience to override our inner sense, it begs the question why we would ever form erroneous second order beliefs. When it comes to the analogous case of proprioception of phantom pains, we have an explanation, but we lack one for beliefs.

2.2.2 Nichols and Stich’s version

Shaun Nichols and Stephen P Stich present their version of the inner sense theory in their book “Mindreading: An Integrated Account of Pretence, Self-Awareness and Understanding Other Minds” (2003, pp. 150-199). Unlike both Armstrong and Goldman, Nichols and Stich only focus on one kind of mental state, belief, and suggest generalizing from this example to other kinds of mental

states. In their version, our beliefs are stored in a “belief box”. Each person has a “monitoring mechanism” which creates second order beliefs by copying a belief in the belief box, embedding it in a “representational schema”, and stores the new belief again in the belief box. The “representational schema” is of the form “I believe that __” (Nichols & Stich, 2003, p. 160). If the belief copied from the belief box is *p*, the new belief to be stored is “I believe that p”.

It is not apparent how literally we are to understand the terms “belief box” and “monitoring mechanism”. One single belief box or monitoring mechanism part of the brain seems unlikely, since no part of the brain seems to corresponds to such a function; a plethora of monitoring mechanisms and belief boxes, on the other hand, scattered all over the brain, begs the question how these relate to each other to create a single, believing mind (Carruthers, 2011, p. 195). Understanding the expression “belief box” this way might be too literal. However, if we are to think of their version as a model of the brain, and we should understand the terms as signifying something less than a concrete entity in the brain, then it is less obvious what this theory actually claims about the workings of the brain, and hence what kind of predictions can be made from it.

These potential problems disregarded, one might note the strong resemblance to Armstrong’s theory. The monitoring mechanism appears to be quite close to the self-scanner suggested by Armstrong, and while Armstrong does not suggest a belief box, it is not obviously incompatible with his account. For the present purpose, the main difference between their accounts can be found in Nichols and Stich’s distinction between “reasoning” and “detecting”, a distinction which Armstrong does not make. The capacity to use information about one’s own or another person’s mental states to make predictions about a person’s behaviour and past and future mental states is “reasoning”. The capacity to attribute current mental states to a person is “detecting” (Nichols & Stich, 2003, p. 151). Only first person-detecting is done by the monitoring mechanism. When it comes to reasoning about beliefs, both our own and anyone else’s, as well as “detecting” beliefs in other people, this is not done by the monitoring mechanism (Nichols & Stich, 2003, p. 161). For this we instead use the ”theory of mind information”. This is a theory – at least in a wide sense of

the word – we use to draw inferences about other minds and mental states using a wide variety of sources, and is more or less what *theory*-theorists assume. This means that the main difference between the theory-theory and Nichols and Stich's account, as they describe it, is that the former takes the ”theory of mind information” to do first-person detecting as well, while the latter suggest a separate system for this.

The main reason for supposing a monitoring mechanism, according to Nichols and Stich, is that there is empirical evidence of situations where people are able to "detect" in themselves, without being able to "reason", and vice versa. In support of this Nichols and Stich refer to studies of people with autism and people with schizophrenia as examples (Nichols & Stich, 2003, pp. 183-192). Persons with autism appear to have problems with third person belief ascription, while being able to do first person belief ascription, they claim. For people with schizophrenia, the case is reversed, they state, and refer to studies where the schizophrenic test subjects could not accurately state their own mental states, while they still were able to do "third person mindreading" (Nichols & Stich, 2003, p. 190). They claim that these phenomena would be impossible if the "detecting" and "reasoning" functions were handled by the same process, which is by the ”theory of mind information".

There has to be a connection between the beliefs in the belief box and the "reasoning" done by the ”theory of mind information". When we predict our own future behaviour, for instance, we rely on our beliefs and this must be captured by the account. Normally, the monitoring mechanism is creating second order beliefs for the ”theory of mind information" to use in "reasoning". However, Nichols and Stich say that the ”theory of mind information" is capable of creating second order beliefs about both oneself and others, and that when this "reasoning" occurs, the ”theory of mind information" is not able to discern what comes from the monitoring mechanism and what comes from "reasoning" done in the past (Nichols & Stich, 2003, pp. 162, 169). The first-person second-order beliefs created by "reasoning" can conflict with those created by "detection", and they do not state how such a conflict might be resolved (Nichols & Stich, 2003, p. 163).

2.2.3 Goldman's version

In "Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading" Alvin I Goldman argues against Nichols and Stich's monitoring mechanism-account and proposes his own version (Goldman, 2006, pp. 223-254). Goldman notes that the monitoring mechanism is not sensitive to different kinds of mental states, and that Nichols and Stich claim that there are separate mechanisms for beliefs, desires and perceptions. Instead of this large number of distinct monitoring mechanisms, Goldman suggests that this is done by *introspection.* He pictures, as Armstrong does, introspection as a non-inferential "looking inwards", modelled on other perceptual processes such as seeing (Goldman, 2006, p. 225).

While Goldman discusses several possibilities, he seems to favour a model that can be described thus: The only mental states that can be introspected are occurrent states, and hence any standing state needs to be "activated" before being introspected. He does not discuss how this is done. When this state is introspected the type and strength of the state is first characterized, supposedly being something similar to "small pain" or "strong belief". The content of the state is then usually *redeployed*, a term defined as "manipulation and repositioning of a piece of mental syntax" (Goldman, 2006, p. 239). If a person has an occurrent state to believe strongly that *p*, introspection first characterizes the type and strength of the state, and then *redeploy* the proposition "*p*". Since this is the last step of the *introspection* process in Goldman's version, it seems plausible to assume that the redeployment involves something similar to Nichols and Stich's "embedding in a representational schema", meaning that a person after introspection has a second order belief.

Sometimes, however, Goldman writes, the mental state is not in a form that can be redeployed. Something a person sees has a "visual representation" in the brain, using a "visual code" (Goldman, 2006, p. 254). A second-order belief presumably does not use this "visual code", according to Goldman, and therefore the "visual representation" cannot be redeployed until it has been translated from one "mental code" to another. This translation occurs, for example, when we self-ascribe beliefs about our current conscious visual experience – *I believe there to be a gorilla on*

the basketball court – but presumably not when we self-ascribe moral opinions – *I believe prostitution to be morally reprehensible*. It is important to note, though, that translation does not involve any inferences in Goldman's account.

The bulk of Goldman's book is devoted to the "simulation theory", which he introduces to explain third person belief-attribution. To do this, we let our minds operate in a "pretence" mode, where we create a belief we do not have and use knowledge of how our own mind works to understand other minds, or our own mind at a different time. Goldman concedes that there are times when we do confabulate and give false first-person belief ascriptions. This is done by a "backup method", capable of both first- and third-person belief-ascriptions. This method is used when introspection is not possible (Goldman, 2006, p. 232). However, Goldman is careful to point out that because this method is used in some cases, it does not mean that it is the only one or the one that we generally use. It would be problematic for his theory if we appear to use this "backup method" every time we happen to test a person's self-knowledge, so Goldman needs to spell out when and why we can expect introspection to fail.

2.3 The theory-theory

Some fundamental aspects of the *theory*-theory have already been stated. The general idea is that we do not have a special method, such as an inner sense, to determine what beliefs we have, or to form second-order beliefs. Instead we make first-person belief ascriptions by the same method we use to make third-person belief ascriptions. That is by making inferences from clues, such as behaviour, using a theory of intentionality, and this is the *theory* of the *theory*-theory. There has been substantial discussion whether or not this actually is a *theory* in the scientific sense of the word – a set of relations or laws built upon, and that could be defeated by, empirical evidence – or is something vaguer, such as a stance or a model, which might not work in the same way as a scientific theory (Gopnik, 1993, pp. 2-3, see also Stich & Nichols, 1998). This distinction, however, has no consequences for the present topic, so "theory" will be used as a placeholder term.

Several philosophers endorse some version or aspect of the theory-theory. One of the more prominent proponents is Alison Gopnik. She bases her view on studies that indicate that children learn how to attribute intentional mental states to others as they grow, and that this ability is parallel to their ability to correctly attribute intentional mental states to themselves (Gopnik, 1993). During childhood, this theory is created and constantly improved upon, which makes it possible to understand what other people believe, hope, fear etc. – and this is the same theory we use when we want to know what we ourselves believe. While it might appear that we have an inner sense that is "reading" our mental states, this is an illusion that stems from our extensive knowledge and experience of this theory of intentionality. The illusion is similar to how an experienced chess player "sees" the right move on a chess board, without having the experience of applying the theory of chess to the situation (Schwitzgebel, 2011, p. 42).

The theory-theory predicts confabulation and introspective errors, which speaks in favour of this theory when considering evidence that people makes introspective errors. A person without direct introspective access to her mental states has to make an inference from the available evidence to determine her mental state, and if the evidence is misleading, the resulting second order belief will be false. While all inner sense-theories claim that introspection can fail sometimes, the theory-theory suggests that such cases can be very common depending on the evidence provided. The theory-theory also says that there is no fundamental difference between the process behind confabulation and behind veridical reports of belief: in both cases the subject employs a theory inferring that she has a certain belief.

One might note, however, that the theory-theory is compatible with the claim that we are *better* at inferring our own mental states, than inferring mental states of others, because we have more information to base these inferences on. While we have no direct access to the belief itself, as we would through an inner sense, we have direct access to our body and can notice the changes of heart pulse, feelings of disgust or perspiration, which usually is not available to other people. However, it is still a matter of *inference* from this evidence when we ascribe beliefs to ourselves.

We have no direct access to the mental state creating the bodily reactions that we use for the inference. One could argue that it is unlikely that we have intentional mental states, or at least mental states as salient as beliefs, of which we are only indirectly aware. Could it really be that I learn what I think of prostitution, for example, from subconsciously perceiving my heart-rate? After all we are directly aware of many things, so there needs to be a reason why some mental states would not be directly available to us. This explanation is not provided within the framework of the theory-theory.

3. Choice Blindness

3.1 Background

Inner sense theories, in all versions, claim that we have non-inferential access to our mental states through introspection. Having this special method to access our own mental states is usually thought to mean that we have *better* access than anyone else to our own mental states. One way to argue against an inner sense theory is to show that we are not more reliable than other people in determining our own mental states. This could be shown if we make errors when self-ascribing beliefs in the same fashion as third-party observers do when observing us.

This might seem unlikely to happen. However, in 1977 Richard E Nisbett and Timothy D Wilson claimed to have found a number of such cases in their article “Telling More Than We Can Know: Verbal Reports on Mental Processes”. In the studies that they refer to, the test subjects were asked to perform a task and then to report on the thought process behind their actions. In the perhaps most famous case people were asked to choose which ones they preferred out of four pairs of nylon stockings. The test subjects were then asked to report on their thought process behind their choice – i.e. why they preferred the chosen stockings. The test subjects gave answers citing properties one might expect to influence such a choice, for instance texture and material quality. The four stockings, however, were in reality identical, and their choice behaviour revealed a preference for socks positioned to the right on the table. Since there were no qualitative differences

between the stockings, and since there was an alternative explanation, Nisbett and Wilson concluded that the thought processes that the test subjects reported were confabulations.

In another study that they cite, test subjects were to connect two cords hanging from the ceiling. The solution, which was far from obvious in the given situation, was to make the chords swing towards each other. The test subjects reported that the solution came because “it was the only solution left”, “the idea of the swinging cord came complete” and one person even reported having being inspired by thinking of monkeys swinging in trees (Nisbett & Wilson, 1977, p. 241). However in all of these cases, the experimenter had made the cord swing, seemingly by accident. Since none of the test subjects reported that as the cause for them finding the solution, it would seem that an independent observer, in this case, would have had an even better knowledge of the actual thought processes involved than the test subject herself. The conclusion that Nisbett and Wilson drew from these and the other studies was that we do not use introspection when reporting on our thought processes, instead we use “a priori causal theories”, and that these are the same that we use when explaining other people’s behaviour (Nisbett & Wilson, 1977, pp. 248-249).

Now, this is not necessarily very problematic for an inner sense account. While we report accurately on our mental states using our inner sense, this sense cannot access thought *processes*. Such an account can say that while it can determine different states at different times, it cannot determine if the first state *caused* the other. This is the explanation Goldman favours. We are, as a rule, not aware of the causes of our behaviour, and especially not *as* causes (Goldman, 2006, p. 232). Determining what caused a behaviour is done by the backup-method that Goldman proposes. Nichols and Stich frame their explanation in terms of their distinction between “reasoning” and “detecting”. Explaining your thought process is “reasoning”. This is done by the ”theory of mind information”, which we apply both in first and third person. We make mistakes when this theory is not sufficient, and Nichols and Stich mean that the results of these experiments are not problematic for their inner sense account (Nichols & Stich, 2003, p. 162).

One might wonder, though, what beliefs the test subjects in the nylon stockings-experiment actually had. If they had the belief that the stockings were different, a false belief, that belief should have been the reason for their report. Their apparent confabulation would then simply have been correct reports, they believed that the stockings were different and reported this. If they had the belief that the stockings were the same, we have two options. Either the monitoring mechanism did not use this belief to create a second-order belief – which would need an explanation – or it did, but the "reasoning" also generated a belief, and these conflicted. This would again call into question how conflicts between second-order beliefs are resolved. Without answers to these questions we do not seem to have any way to tell if our beliefs ever are genuine or not, and the explanation appears *ad-hoc*.

While the "choice blindness"-experiment that are discussed in this essay are reminiscent of the experiments mentioned by Nisbett and Wilson, they are also an extension of another type of experiments. In "change blindness" and "in-attentional blindness"-experiments it has been shown that we are not as aware of changes and objects present in our visual field as we might think. When, in one experiment, test subjects are instructed to count the number of passes of a basketball between players, fifty eight percent of them do not notice a person in a gorilla suit walking across the court, even though it is clearly a visual and supposedly highly unusual experience (Simon & Chabris, 1999). In other experiments people have been shown not to report differences when the colour of the background is changed, or the upper-lower case of lettering is changed, to mention just a few examples (Dretske, 2004, p. 5).

The first choice blindness experiment has many similarities to both "change blindness" experiments and the experiments described by Nesbitt and Wilson. In this test the test subjects are to choose which of two faces they prefer (Johansson, et al., 2005). They are then allowed to look at the face again and are asked to state what made them prefer this face. This is then repeated with several pairings. However, in some cases, some of the faces they are shown are in reality the ones they did *not* choose. Test subjects do not generally notice this, and when they state their reasons

behind their choice, they instead refer to features of the face they are currently looking at rather than the face they actually chose. Similar experiments with the taste of jam and the smell of tea have produced similar results (Hall, et al., 2010). This might be surprising – how could we not notice that we are eating mango jam instead of apple and cinnamon jam? Goldman discusses this and considers the possibility that while the test subjects are supposed to answer questions about the old picture, they are actually forming new opinions about the current picture, and that the details of the old picture have faded in their memories (Goldman, 2006, p. 235). These results, however, have no clear new implications for the inner sense theories, other than adding "reasons for our preferences" to the list of things we cannot use our inner sense to find out. It seems plausible that we do not have deeply rooted preferences between faces of strangers. We know that our senses are – surprisingly – fallible, so if we do not notice a person in a gorilla suit clearly visible in front of us, it does not seem unlikely that we would not notice the difference in taste between mango and apple either. Preferences for jams are, at least for most people, not deeply rooted and the experiment appears to be more of a gustatory equivalent to an optical illusion. Later choice blindness experiments present more serious challenges to the inner sense theories, and they are the ones that will be the main focus of this essay.

3.2 The experiments

Many would agree that preferences for stockings, faces, jam and tea are not deeply rooted in our identity. While we probably would state such a preference if we were asked, most would say that it does not matter very much. In the following two experiments the test subjects were asked about things most people *do* believe matter: moral issues and politics.

In the first experiment, "Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey" (2012) Lars Hall, Petter Johansson and Thomas Strandberg asked participants to state on a nine point bidirectional scale their level of agreement with statements concerning either moral principles (test type 1) or hotly debated moral topics (test type 2). The scale went from 1: completely agree, to 9: completely disagree, and 5 meant "neutral"

or “undecided”. In this test 160 participants were randomly recruited in a park as to have a neutral and relaxed environment.

In the second experiment, “How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions” (Hall, et al., 2013) the topic was political issues. The statements dealt with issues where the two main political coalitions in Sweden hold opposite positions, and the participants were to mark their agreement to the statements on a 100 mm scale, from completely agree to completely disagree, where the middle represented “neutral” or “undecided”. The survey was presented as an “election compass”, which would indicate which coalition you should vote for on the basis of your position on the different issues. 162 participants were recruited in various places in Lund and Malmö. Before starting the test, participants were asked to state their voting intention and certainty by marking on a scale ranging from one of the coalitions to the other.

3.2.1 The statement phase

In the first part of the experiments, which I will refer to as “the statement phase”, the test subject report on their beliefs with regard to different issues. In type 1 of the first experiment a statement could be: “It is more important for a society to protect the personal integrity of its citizens than to promote their welfare”, or “Even if an action might harm the innocent, it might still be morally permissible to perform it”. In type 2 of the first experiment a statement could be: “The violence Israel used in the conflict with Hamas is morally defensible despite the civilian casualties suffered by the Palestinians”, or “It is morally defensible to purchase sexual services in democratic societies where prostitution is legal and regulated by the government”. They were also asked to rate how strong their moral opinions in general were and if they were politically active or not. In the second experiment a statement could be “Gasoline taxes should be increased” or “Healthcare benefits should be limited”.[1]

[1] The surveys were given in Swedish. The translations are from the cited articles.

3.2.2 The deception

In the first experiment the deception was executed by a self-adhering piece of paper that covered original statements, and was attached surreptitiously in the following way. The questionnaire consisted of two pages stuck to a clipboard. The test-statements were on the first page, and the second page contained questions about age, gender, etc. Stuck to the back of the clipboard was a paper with almost the same statements as on the first page of the questionnaire. This paper had a self-adhering backside, so that when the test-subject turned the first page over, the paper adhered itself to the first page, over the original statements. In this way, the original statements were changed without the test subject's knowledge. Some, but not all statements were manipulated.

The statements were changed to, for example, "It is more important for a society to promote the welfare of its citizens than to protect their personal integrity" and "If an action might harm the innocent, then it is not morally permissible to perform it" for test type 1. Of test type 2, the statements now read: "The violence Israel used in the conflict with Hamas is morally reprehensible because of the civilian casualties suffered by the Palestinians" and "It is morally reprehensible to purchase sexual services in democratic societies where prostitution is legal and regulated by the government". Only the statements were covered by the self-adhering paper. This meant that their markings now represented the directly opposite view in the manipulated experiment.

In the second experiment, the deception was made in a slightly different way. When discussing the result of the first experiment, the choice blindness group proposed that it was easier for people to detect complete reversal of opinions, and that the number of successful manipulations might increase if the level of agreement was altered in a more believable way. In the second experiment, the experimenter looked over the shoulder of the test subject while she was filling out the election compass and filled out an identical answer sheet. The experimenter copied some of the answers and changed some, so the compass would show the person securely belonging to the other political coalition compared to their original statement. In this version the experimenter also tried to create a "believable opinion profile", so that the manipulations were less apparent and the beliefs were more

coherent. This manipulated answer sheet was slipped over the original without the test subject's knowledge. In both experiments non-manipulated control groups were used.

3.2.3 The reviewing phase and the debriefing phase

In the second part of the experiments, which I will refer to as "the reviewing phase", the test subjects were asked to look again at their answers, which now were surreptitiously manipulated. They were asked to read aloud some randomly selected statements, discuss their stated agreement with them and, if they wanted to, change their answer. If they during the reviewing phase changed any manipulated answer, the trial was considered to be "spontaneously detected". For the second experiment this meant that the election compass would change back to their real coalition preference if a sufficient number of the manipulations were detected. Only one test subject suspected being deceived. The rest who detected the manipulation said that they must have made a mistake when answering, such as misreading or not understanding the statement.

This was then followed by what I will call "the debriefing phase". The test subjects were now asked a series of questions. The first questions were about what they thought in general about the questionnaire and they were then asked if they had thought anything at all to be odd or strange during the test. Finally, they were shown how the manipulations were done, and asked if they had felt any kind of suspicion, or anything at all indicating that something was wrong. If they at any point during the debriefing phase indicated that this was the case, the trial was considered to be "retrospectively detected". In the following sections I will sometimes use the phrase "fail the choice blindness-tests" to mean being deceived by the test, and to fail to notice the deceptions and make corrections.

3.3 Results

In the first experiment concerning moral principles and moral issues, a total of 47.2 % of the trials were detected; 41.3 % spontaneously and 4.4 % retrospectively. In this case, each manipulated question counts as a trial, and hence the total number of trials was 320 (Hall, et al., 2012, pp. 4-5). In the second experiment 53 % corrected one to four manipulated answers. Despite these

corrections, 92 % had their political coalition shifted compared to their original statement, and still endorsed this result as saying something about themselves (Hall, et al., 2013, p. 3). The number or percentage of spontaneous and retrospective correction was not reported in the article.

In the first experiment, the discussion during the reviewing phase was recorded and transcribed. This transcription was then shown to three blinded researchers. These were asked to read the transcription and state what level of agreement they believed test-subject had marked on the scale. These independent researchers were able to predict the level of agreement with the statement with high accuracy for both the manipulated and for the non-manipulated trials (Hall, et al., 2012, p. 5). The research group have in earlier experiments argued that that this method shows that there are no lingering subconscious suspicions noticeable in the test subjects' language use (Johansson, et al., 2005).

There were no correlations between detection and age, gender, amount of time spent on the questionnaire or self-rated strength of moral convictions in the first experiment. There was a correlation between detection and level of agreement with a statement, meaning for example that if a test subject had marked an agreement of 1 for the statement, the subject was more likely to notice that something was wrong, most likely because of the drastic change. However, 31.4 % of the manipulated statements at either endpoint of the scale were not detected. In the second experiment, there were no correlations between detection and age, gender, level of political engagement or previously stated political affiliation. Since the manipulations were done differently, there were no extreme manipulations from one endpoint of the scale to the other. It might also be noted that most test subjects reported surprise when realizing that they had argued for another opinion than the one they had stated at the beginning.

3.4 Implications for the inner sense theories

It seems plausible to assume that our beliefs about moral principles, moral issues and political issues are both stronger and more stable than our beliefs connected to our preferences for certain face types or kinds of jam. Furthermore, when we answer questions about the former, we do appear

to do some kind of introspection, to carefully look inwards before reporting what we think. Petter Johansson argues in his dissertation that this kind of test shows that inner sense theories are wrong; "If we are supposed to know our minds from the inside, we should know why we do what we do" (2006, p. 20).[2] The reason why we believe ourselves to have privileged access to our beliefs is that no one is usually able to prove us wrong, and as observers we have to take their word for it. However, in these experiments this is not the case, and Johansson argues that the results indicate that theories such as Goldman's are flawed (2006, pp. 20-21).

What does Johansson need to postulate about the workings of the inner sense so that it conflicts with the choice blindness results? The following is a description of what an inner-sense theory could say happens during the test; in the discussion-part I will look at alternative ways to describe what happens during this process. All three inner sense theories agree that when we are asked to state our agreement we survey our beliefs, either with the self-scanner, the monitoring mechanism or using introspection. We then form a second-order belief which we express using the scale provided. This much is uncontroversial. What happens in the *reviewing phase* is less so. None of the theories describe how we detect having made a mistake when expressing a belief, or how we detect errors in general for that matter. I suggest that we need to match the combination of the statement written and our stated agreement with our own belief for us to be able to detect the manipulation. It is obvious that we cannot notice the manipulation if we do not compare the statement with our belief. In the reviewing phase, we first read the statement and our stated level of agreement out loud. We might match the statement and the belief automatically at this point: we again introspect our belief with regard to the statement and compare our beliefs with the one that is stated on the paper. In order for this to happen, we also need to form a corresponding belief of what is said on the paper to compare with our own belief. In manipulated conditions these are conflicting, and we correct the one on the paper. The manipulation should, in this interpretation of the inner

[2] This way of phrasing the challenge to the inner sense theories differs from what I take it to be. The inner sense-theories do not claim that we are aware of the thought-processes behind our actions.

sense theory, (almost) always be spontaneously detected, so it is problematic for the inner sense view that this does not happen.

We are then asked to discuss our view. Now, while Armstrong is silent on the matter, both Nichols/Stich and Goldman claim that this is not primarily done by the inner sense. Instead, this is done by the theory of mind information or simulation/backup system, respectively. I interpret this reasoning-faculty as the capacity to use beliefs we have detected by using the inner sense, to form chains of reasoning. As an example the reasoning faculty might explain my belief that prostitution is reprehensible by stringing together the belief that human beings are not objects, and that I believe that only objects should be sold. This should be possible as long as we have some beliefs that work as a part of a chain of reasoning that forms a permissible output – that we have some beliefs which are suited to be justification for the belief in question. Our ability to reason about a belief we do not have is therefore compatible with the inner sense theories. However, it seems reasonable to claim that a reasoning faculty should use the beliefs that we have, and hence we should at least be less able to discuss a belief we do not have, since we should have fewer beliefs suitable for this discussion. This is another problem for the inner sense theories, since we are every bit as able to give arguments for the beliefs that we did not state. Furthermore, this discussion would probably make it more likely that we use our inner sense again with regard to our beliefs concerning the statement, which would cause a matching to take place.

Finally we are asked questions about the test itself. If we have formed any beliefs about the test being strange, we would of course use our inner senses to self-attribute and report on these. However, if we have not detected anything being strange, there is nothing to report on.

With this interpretation of the inner sense theory, we should not fail the choice blindness test if the inner sense theory is correct. If we truly use our inner sense to access and self-ascribe our beliefs when we state them, we should use the inner sense again in the reviewing phase and notice the manipulation. Since this does not happen and we are indeed deceived, the inner sense theory could not be correct. Furthermore, the *theory*-theory gains support from the choice blindness

experiments. The theory-theory states that we use clues in the environment to determine our own beliefs. A paper showing our own statement is of course such a clue, and one we are inclined to trust. Markings on papers do not change themselves, and we trust the experimenter. The question is now: is there an interpretation of the inner sense theory where it does not conflict with the choice blindness experiments, and how plausible is such an interpretation?

4. Discussion

The easiest way for the inner sense-theory to escape these problems is to say that most people *did* notice the deception, but were too embarrassed to say anything or, alternatively, were *subconsciously* too embarrassed to say anything. “Embarrassed” might be substituted for any kind of emotion that we can think would suppress either a detection or a report. This does not seem to be a good explanation. Taking in the sheer number of people who were deceived, and considering the surprise most expressed when being told about the deception, together with the analyses of the verbal reports, there is nothing to indicate that people were too embarrassed to say that they noticed the deception. Furthermore, to claim – factually or otherwise – that you noticed something strange during the test, and did not say so, would seem as an easy way to “save face” in the debriefing phase, and very few appear to have done this since only 4 % of the test subjects in the first test reported retrospective detection. I believe the test setup provides sufficient reason to believe that this was not a major factor behind the results and this line of defence will not be investigated further.

There are other, and better, ways to defend the inner sense theories, and I will present two such ways. The first way (discussed in section 4.1) is to say that the inner sense does not work in the way described in section 3.4. There are several points in the test procedure where the inner sense could behave differently from what I believe Johansson needs to assume. The main point of this defence is that introspection sometimes, for certain beliefs, is harder to do than for other beliefs.

The second way is to note that the role deception plays in these tests is problematic. The fact that we can be deceived does not allow us to draw as many conclusions as it might seem.

However, when looking at the wider picture, the inner sense theories still faces problems. I see two main problems (discussed in section 4.3). The first problem is that it is not enough to be able to explain away evidence such as that provided by the choice blindness experiments. If there are no testable implications of the inner sense theory, it will be *ad-hoc*. I will later briefly suggest three ways to test the inner sense theory, if it works in a way that cannot be tested by the choice blindness method. The second problem is that the overall evidence still points in favour of the theory-theory. Constructing a test that we would not fail if we did have an inner sense, as previously described, is not enough. One would also need a test that is decisive between the theories, i.e. that we would fail if the theory-theory was true and not fail if the inner sense theory was true, or vice versa. The improved tests that I suggest are, in my view, still not capable of this and this area is ripe for development.

4.1 Clues and cognitive economy

It might be that the strange results in the choice blindness experiments does not show that there is no such thing as an inner sense, because of some peculiarity of the test setup. I will suggest three possible lines of defence. The first line of defence is to say that the matching between the statement on the paper and the subject's belief does not happen automatically, and that if it did, we would not self-ascribe contradictory beliefs. The second line of defence is to say that the means by which the choice blindness group tries to make us use our inner sense and do the matching is insufficient. The third line of defence is to say that matching does happen, but that we never become aware of this.

In section 3.4, I described why we would not fail the choice blindness-tests if we had an inner sense. The main point in this description is the *matching*. I imagine the matching process as a comparison between the statement on the paper and our beliefs, and I believe it to be important for the following reason. We can think of this as two separate tasks: a/ arguing for *your own* belief regarding the morality of prostitution, and b/ arguing for *the belief on the paper* regarding

prostitution. In non-manipulated conditions, these tasks are the same. The key point in noticing the deception is realizing that these tasks are suddenly *not* the same. For this a subject needs to understand the statement stated on the paper, and then compare that statement with her own belief.

4.1.1 The first line of defence

The first step in the first line of defence of an inner sense theory saying that this is not trivial and not something that happens automatically. To me this idea has some intuitive support; sometimes I can read an editorial in a paper and instantly know that I agree with or oppose the main point, sometimes I just read, and it is only when I am asked if I agree or not that I consider this and give an answer. The test subjects are certainly asked “So you agree with this statement?”; however, it could be seen as more of a rhetorical question than a genuine question, and not something that makes us consider our position once again. It could be that unless we are clearly and directly instructed to consider our position once more, we take any question of that sort to simply be a request to make us talk more.

The second step is to propose it being *harder to* use the inner sense in this case. Now, the inner sense is supposed to provide direct and non-inferential access to our beliefs, and this is not supposed to be hard. It should be like looking in a bag and answering what things there are in the bag: it should be effortless and instant. An inner sense theory could claim that this idea is a bit naïve. The belief which we self-ascribe is an answer to a question, and it does not seem plausible that we have beliefs ready to be used as answers to any question imaginable. It seems to me as if the theories need to allow for us to sometimes do a “transformation” or an “aggregation” between the beliefs that we have and the requested output, and that without such a processes the inner sense theory also need to address how we can answer questions that we have never heard before. By transformation, I imagine a process which changes the propositional content of a belief that is very similar to the one required to answer the question. Perhaps I believe that *people* who prostitute themselves are morally reprehensible, while the statement concerns the act of prostitution, concerning which I had no belief. I might then transform this first belief to suit the question. By

aggregation I imagine the case where we have several beliefs that are jointly sufficient to answer the question. These beliefs are aggregated into a permissible output.[3]

Depending on what beliefs you have, answering a question about serious matters, such as prostitution or the Israel-Palestine conflict, might in contrast be very complicated. This means an inner sense theory could stop right here, and say that this is the whole problem. It might be that transformation from the same belief can render different results. The reason why we do not notice the manipulation is that when we introspect our beliefs again to do the matching, the transformation or aggregation process gives an answer, i.e. a belief ascription, which this time happens to be compatible with the statement on the paper. This could be the answer in some cases, however, the sheer number of successful manipulations does not lend much credibility to this solution. For the idea of transformation or aggregation to be at all plausible, the processes should result in the same, or almost the same belief every time. If we were trying to explain how people could have failed to notice a shift from a four to a three, this could be the reason. Many people who gave highly polarized answers were fooled, and we would need a separate explanation for those cases.

Instead, a better way to defend the inner sense theory is to claim that introspection, while direct, is hard in the sense that it is cognitively taxing. Compare it with the act of remembering, and answering the question "What did you have for dinner the Tuesday before last?". To me, and I believe to most people, this is a question that is a bit hard to answer, and can take some time. I might have to consider what I did that day, at what time I got home or if I went by the supermarket. However, I usually end up with a distinct answer that I am certain of, a memory of what I ate that night. It might also be the case that I do not remember what I ate, and I do not arrive at a memory of that night. The point here, is that even when it takes me a surprising amount of time and effort to

[3] One might question if aggregation/transformation is not really an inference, which would mean that the inner sense's access *is* inferential and would perhaps be a reason to exclude a process such as aggregation/transformation. On the other hand, something is needed to answer how we can answer questions to which we have no directly corresponding beliefs. This also raises the larger issue of how beliefs are formed and individuated – if the transformation/aggregation is creating a new belief, and if this happens every time we are asked a slightly different question. Such a discussion lies beyond the scope of the essay, and I will just assume that transformation and aggregation does not require a *theory*, such as the one theory-theory suggest, and that this is not a case of inference in the proper sense.

find the memory, when I actually do it is not because I inferred the answer, it is because I remembered it. The memory is there for me to access directly, it is just a bit "buried below" other memories. Trying to remember what you ate more than a week ago is cognitively taxing.

Using the inner sense might be cognitively taxing in much the same way. There are several studies, not least in behavioural economics, which show that we shun cognitively taxing tasks if we can, and instead use other, simpler, methods of answering questions.[4] A proponent of the inner sense theory might then answer in the following way: of course we do, in ideal circumstances, introspect the second time as well. But since we are lazy, we only do this when we have strong reasons to, as for instance if we begin to suspect something is wrong. Since the manipulations were skilfully done, we never become suspicious, we do not use introspection again, and we never detect the manipulation.

A reasonable objection to this is: why would the belief be so hard to introspect in the reviewing phase. Sure it might be "in the bottom of the belief-pile" during the statement phase, but when we have introspected it once, it should be easy to introspect again, given that we formed a second-order belief during the statement phase. This is indeed a problem for the inner sense theory. One answer could be that it simply does not work that way – a belief that is hard to introspect remains hard to introspect, at least until you have "introspected it" a number of times – even for the beliefs that do not need transformation or aggregation. This warrants an explanation as well, and hence this first line of defence is in need of support on many points.

4.1.2 The second line of defence

The second line of defence is to say that the means by which the choice blindness group tries to make us introspect and do the matching are insufficient. This matching process needs to be kick-started in some way. What is perhaps most remarkable about the choice blindness-experiments is not that we do not notice the deception right away, it is rather that we are able to *discuss* and *explain* answers we did not give without realizing that we did not give that answer. We discuss it as

[4] See for instance Tversky & Kahneman, 1974.

if it *were* our answer. Furthermore, we appear to be *as good* in discussing positions we do not hold as positions we do hold. Now, as noted in section 3.4, this may not be impossible to explain with the inner sense theory. Nichols and Stich, for instance, say that "reasoning" about a belief is done by another system, the theory of mind information (Nichols & Stich, 2003, p. 161). Goldman claims that simulation of a target's mental state can be highly inaccurate, which in this case would mean that we give our reasons after a simulation based on a false "pretence-belief" about ourselves (Goldman, 2006, p. 39).

A first version of this second line of defence is to say that the discussion does not kick-start the matching process, since it does not involve the inner sense at all. This statement needs to be moderated. Of course the "reasoning" has to involve the inner sense in some way, since the discussion ought to be based on at least some of our beliefs, otherwise we confabulate all the reasons too. Rather, they would say that reasoning about a belief does not necessarily involve introspecting that very belief. The "theory of mind information", as an example, is working to answer the question "why do I think *this*?" or "why *would* I think this?" but without considering the question "by the way, do I really think *this*?". The system is insensitive to the origin of the self-ascribed belief, and does not notice that it was not produced by the monitoring mechanism.

Is this reasonable? The first objection to this second line of defence might be to say that this amount of introspection should make us notice that the belief in question was absent. To this, one could simply reply that it does not work that way – introspection works on beliefs that actually "are there" and does not make us notice an absent belief. This begs the question how we realize that we are unable to answer a question, because we have no beliefs on the matter – when we are asked about something we do not know anything about.

The second objection would be that when we are asked to explain a position we do not hold, there would not be any suitable beliefs for the "backup-system" or the "theory of mind information" to use. However, this is not satisfactory. Even people with quite polarized opinions typically have beliefs that could be used to discuss a different position from their own. For instance, even people

who think that prostitution is extremely reprehensible also usually believes that one has a right to control one's own body, which could be used by the reasoning-faculty.

The third objection is a variation of the second one: Since we have (at least mainly) coherent and consistent beliefs, there should be fewer beliefs compatible with a position we do not have. Because of this, we should be less able to discuss a position we do not have. This seems plausible – however, it might still be the case that we are *less* able, since the interviews were rather short. Perhaps we have enough compatible beliefs to keep up a discussion about a position we do not hold for a few minutes. After all, these issues are very complex, and most people know of both pro and con arguments in these issues. For instance, even someone who does not agree at all with Israel's actions against the Palestinian people, understand that this is a complex issue and that Israel's status as a nation is constantly being questioned, and that anti-Semitism is strong in many places. Holding such beliefs would presumably facilitate a couple of minutes of discussion. Perhaps the telling difference in the quality of the discussion shows itself only after a longer period of time, such as an hour.

This means that it is not a very strong counter argument to the second line of defence to say that the discussion in the reviewing phase should kick-start the matching process because we should be *less* able to discuss a position that we do not hold. Of course, it still appears strange to think that we would be able to argue at all for something that we do not believe without realising this. However, this follows from the inner sense being postulated as a separate entity. One might have a naïve idea that we have chains of connected beliefs which provides justification of the belief in question, an ordered logical structure which we report on when we explain our beliefs. I do not think that beliefs are generally formed in this way – as a result of a logical deduction from other things we believe, but if they are formed this way, then this result is of course even more incredible. However, the inner sense theories admit that we are bad at reporting on our cognitive processes, and are bad at knowing that the causes for our beliefs are the actual causes, so this is not a specific

problem for these tests. As such, this second line of defence remains a possible explanation for the inner sense-theory.

4.1.3 The third line of defence

Finally, the third line of defence is to say that the inner sense introspects the original belief, does the matching, but that the conflict is resolved in favour of the belief stated on the paper and there is a change of belief. Nichols and Stich surmise that since the ”theory of mind information” is also capable of producing second-order beliefs, there must be a way to resolve conflicts between beliefs. They do not expand on how this is done, so I will just note that this is also a possibility. We *do* re-scan and match, but we resolve the conflict by abandoning our previous belief in favour of the new one and adopt the written belief as our own.[5]

One might also say, as a variation of this third line of defence, that a matching indeed takes place, however, what is being matched with the statement on the paper is not our introspected belief. Instead it is our memory of our own statement. “Was not my marking over there?” our subconscious might say to itself when faced with the mismatch caused by the manipulation. Knowing that markings on a paper is more reliable than our memories, the memory is discarded as faulty. We could in principle introspect again and do a real matching, which would result in a detection, but refrain from this because of cognitive economy, since the short-term memory is more easily available. Of course, there are problems with this kind of defence. There appears to be no traces of this subconscious process where we match our memory with the statement. While we are not aware of the process at the time, this seems as a memory we would retrieve when we are told that we were deceived. Furthermore, there are plenty of opportunities for this memory to resurface when we are discussing the statement, which also is a kind of activity that sometimes serve to jog our memories, and there is not any apparent explanation for why this does not happen in this case.

[5] Actually, the ”choice blindness” research group performed experiments where the deception was not revealed over longer periods of time (forthcoming). Even several days later, test subjects still reported the manipulated answers, instead of their own original, meaning that they did not revert to their old position.

The debriefing phase does not require its own explanation. If we did not introspect and match, there would never be any possibility for us to feel that something was strange. When asked if we think we could have been deceived this way, we say *no* because we do not have any lingering idea of being fooled, and because we over-estimate our own abilities.

Summarising, one way for an inner sense theory to avoid the troublesome implications of the choice blindness experiments is to claim that introspection is cognitively taxing and that the methods employed to make us use our inner sense are insufficient. Even if the belief is introspected again, for us to notice the deception we also need to match the belief on the paper with our own, and reject the one on the paper. An inner sense theory can also say that the conflict between those beliefs are resolved in favour of the written one. Finally, it is possible the results of the choice blindness-tests might be explained by all of these lines of defence combined, and that errors of different degrees might have different explanations.

4.2 Deception and detection

Another way to avoid the implications of the choice blindness experiments is to question the very foundation: the deception. People often believe they are much less fallible than they are. We are easily caught up by phenomena such as priming, the halo-effect and conflation of cause and effect. When our fallibility is demonstrated to us, we are surprised, because it does not match our self-image. The question, however, is not about being infallible. The inner sense theories all admit for us to be fallible to some extent.

The question is rather: Can we draw any conclusions from the fact that we can be deceived? The analogy with vision might be illuminating: While we are easily susceptible to optical illusions, such as the Müller-Layer illusions where two lines looking to be of a different length because of the arrows at each end pointing in different directions, no one would from this conclude that we do not ordinarily know the length of lines or that we do not have direct access to our visual perceptions. I will argue that the deception in the choice blindness experiments is problematic in much the same way using the following examples.

Suppose I know the way by car from my apartment in Stockholm, Sweden, to the choice blindness laboratories in Lund, Sweden. If I were to go there, I would not go astray, even though I've never actually been there. Now, suppose I am to visit the choice blindness group in Lund, and I mean to go there by car. The choice blindness group offers me a route instruction for the way from my apartment to the research facility, which I bring along even though I do not need it. While they claim that the instructions are a print-out from Google Maps (a map-making company I trust), they have made several changes to it, which leads astray any person following it. If I were to go to the lab and went astray because of the faulty instructions, does it mean that I didn't really know the way to Lund? To me, this seems to be clearly false. Of course, if I was given the instructions and asked to compare them with my own idea, I might notice the deception. The point is that this is a case where I could be deceived, and we still would say I had knowledge.

My second example of deception is from a Swedish radio show. In the 1990's there was a radio show in Sweden, "Hassan", which entertained the listeners by making prank calls. In one episode, one of the hosts made a call to a randomly selected person, introduced himself to the woman answering and said (my transcription and translation):

> Host: "Hello hello, we have found out something."
> Woman: "What?"
> Host: "'Babar' backwards is 'rabarber'."
> Woman: [Pause] "And..?"
> Host: "It's amazing! It's called a palindrome. Words you can say both forwards and backwards."
> Woman: "Yes..."
> Host: "'Esso' is one of those words."
> [Pause]
> Woman: "No..."
> Host: "Yes."
> [Pause]
> Woman: "Osse."
> Host: "Well, almost 'Esso'. But 'Rabarber' we agree on?"
> Woman: "Yes, that is true."
> Host: "Thank you very much."

It seems to me that few would from this conclude that the woman actually did not know how the word 'rabarber' (a word meaning 'rhubarb') is spelled, even after having made sure she understood the idea of "backwards" by using the detected 'Esso/Osse' example. What I suggest is, as I suppose

most people would do, that the woman does not actually consider what the word 'rabarber' is backwards – it is cognitively taxing to spell something and especially to spell something backwards.[6] The very idea that someone would call her and fool her about this is probably to her every bit as absurd as the idea that someone would call her at all regarding this.

To see why the deception aspect of this fails to pick out anything particularly interesting, assume that the host instead had said: “which one of these words is 'rabarber' backwards: 'babar' or 'rebrabar'?”. Surely an equally absurd question, and while the answer really is empirical, does it really seem likely that the woman would have picked the wrong one? The order of the words matter here, of course. It does not seem likely that she would have agreed with the “rabarber/babar” case if she had done “esso/osse” first. However, this could be said as well for the choice blindness tests. The choice blindness group write in an earlier article that when a detection came early, it increased the likelihood of further detection, which most likely holds true for the later experiments as well (Johansson, et al., 2005, p. 117). To go back to the road instruction-example earlier, compare with a case where I was offered two maps and had to choose which one that I should use. Would I pick the wrong one, if I knew the way? Or assume that I realized that the choice blindness group was an unusually tricky bunch, and I started to question their motives – would I still go astray or would I instead start to use my own idea of how to drive there?

Goldman compares an earlier version of the choice blindness-tests, where the participants chose between faces, to the Stroop effect experiment (Goldman, 2006, p. 234). In this task a participant is asked to look at words written in different colours, and is asked to say what colour the words are written in. However, the words on the paper are the words for different colours, and the words are not written in its respective colour. Hence the word “blue” might be written with a green colour. The test subject is faced with the conflict between the word that is written and the colour it is written with. Goldman uses this to show why a person might make a mistake when describing

[6] One might also wonder what the woman thinks a palindrome is; the usual definition of a palindrome is: a word or a sentence which is the same word or sentence when spelled backwards. The definition implied here is: a word which is still a word when spelled backwards.

what was attractive about the face that was chosen. Since the test subject looks at another face, she cannot help herself to use characteristics of the new face when describing what was nice about the old one that she chose, especially since she is unaware that they are not the same. This, Goldman thinks, is comparable to how we in the Stroop test find it hard to say the name of the colour rather than the word.

I am not convinced that the Stroop test and the choice blindness tests are all that similar. However, the Stroop test is an excellent example of the problem with deception. When we are deceived or, as in the Stroop case, when someone is deliberately making it hard for us to perform the task, we make errors. However, the Stroop experiment does not show that we are not as familiar with the colours as we thought or that we have problems naming them. When the words are in the same colour or when the words are unrelated to colour, the task is easy for us. In the same way, we cannot conclude from the choice blindness test that we do not know our own beliefs as well as we thought, or that we do not have direct access to them. We can only make a much more careful conclusion, such that we are more susceptible to deception than we might think. We cannot draw any strong conclusions as to *why* we are so susceptible to deception.

I believe that as soon as some element of suspicion is entered into the experiment, the outcome of the experiment will be different. Imagine that the experimenter had been purposefully absent minded, or dropped the batch of papers, appeared biased, or in some other way been clearly inept at performing the experiment. My intuition is that this would have changed the outcomes of the experiments and increased the detection rate substantially. Petter Johansson agrees with this, and says the experiment was designed in such a fashion that we *would* fail it, and that it would have been easy to construct a version where we would not fail it (personal communication, May 25, 2015). The whole point of the experiment was to show that it was possible to make us do these apparently contradictory belief ascriptions. Our senses are only supposed to work correctly in the right conditions. That we are able to deceive them shows only that they are not infallible. The same

might be said for the inner sense: It works in the right conditions and the conditions of the choice blindness experiments does not qualify.

4.3 Comparing the evidence

It is therefore possible for an inner sense theory to avoid the apparent problematic implications of the choice blindness-experiments. First, it is not clear that the inner sense would have been operating during the review phase of the experiments. Second, the experiment shows only that we can be deceived about what we believe, not that we cannot use our inner sense to access the beliefs. Thus, the inner sense theories might appear to be saved.

However, when looking at the whole picture, things look less optimistic. While the results can be explained away, the theory-theory still has the advantage of *predicting* these results. Furthermore, we want the inner sense theory to be testable. There has to be some test that we *would not* fail if we have an inner sense, and preferably at the same time *would* fail if the theory-theory was correct. All other evidence being equal, we should prefer a theory which has testable consequences.

Such an experiment is not easy to formulate. For instance, one could think that asking the subjects to discuss their statement for as long as they possibly can would solve the problem. If the inner sense is providing beliefs to a "reasoning-faculty" we would expect this discussion to be shorter for beliefs that the person does not have than those that they have, since there presumably are fewer beliefs that can be used as arguments for the opposite belief. This test would, however, have very limited implications. It is only if we are as able to give long discussions about an opposite belief that we have some reason to doubt the inner sense theory. The theory-theory might also claim that we would be less able to have long discussions about a belief that we do not have, since the "theory" that we use is based on the beliefs that we do have.

A possible problem with the choice blindness-experiments discussed in this essay might be that the beliefs concerns very serious matters. Perhaps people who were deceived in the test had an implicit bias towards the opposite position of what they originally stated. So one way to test their

real beliefs is to make participants go through an implicit bias-test before doing the choice blindness test, and see if people with implicit bias more often also fail the "choice blindness"-test.

One could also imagine asking about beliefs in matters where we people do have different and clear opinions, but that we do not feel are very controversial. Are we, for instance, choice blind when it comes to music taste? Would I be able to argue convincingly for the statement "Justin Bieber is the true king of pop" without noticing that someone had tampered with my answer? Could I be deceived about my answer about whether it is permissible to dig into the butter-spread when making sandwiches, or if you should scrape the surface? What topics can we be choice blind about – and what does this say about us, and about our brain? This would, of course, be an investigation in to the choice blindness-phenomena, and not primarily into self-knowledge theories.

Another way to determine if there is an inner sense might be to introduce more ways of detecting subconscious traces of a matching taking place, such as measuring skin conductivity or tracking eye-movement patterns.[7] There might be a matching taking place that never is close enough to our consciousness to verbally express the result. Of course, while the theory-theory says that we do not directly introspect our beliefs, it could allow that we are able to do some version of this matching anyway. After all, several of the trials were detected, and a theory-theory would have to describe how this is done.

This is indicative of the deeper problem with testing these theories empirically. Imagine a different test, where the test-leader drops your paper in the reviewing phase, gathers them up again and hands you the one she thinks is yours. This would presumably introduce enough suspicion for you to go through the answers again and check if they are alright. However, you might simply be *looking at* the test, to see, for instance, if the markings look like they were made by you hand, or if they are in a pattern that you remember. The deception part could still obscure the introspection.

Since deception could be a problem, let us remove it: Imagine an experiment where you state your opinions in several matters on a 100 point bidirectional scale, such as moral dilemmas and

[7] This has been done, by means of measuring pupil dilation, and does not appear to be indicative of subconscious detection (Pärnamets, submitted).

political statements, and perhaps taste in music. The researcher then takes this away and hands you a new sheet where you are to fill out your answers again, however the statements are re-formulated so that they are asking the exact same question but in a different way, they are placed in different order and they are so many that you could not be expected to remember your answers.[8] You would have to use introspection once more, and this may show if we have an inner sense or not.

However, this is not necessarily the case. Imagine the following case: Your scores on the first and second test match really poorly – for two thirds of the questions, we give the complete opposite answer on the second test, like the test subjects in the choice blindness tests. What would such an answer tell us? The theory-theory could interpret this as being the result of us making inferences from clues, and that we inferred the same beliefs for one third of the questions and different beliefs for the rest. This might be because either we had different clues to infer from for some reason, or that we made different conclusions because our *theory* was not specific enough.

On the face of it, such a result (or one with even more different answers) would seem to go against the inner sense theory. After all, since the statements are about the same thing and there is no deception we should access the same beliefs and we should answer almost all the same. However, one might argue that a more reasonable answer to why we scored so badly is that in some of these matters we actually do not *have* a strong belief. We might *think* that we do, because of other psychological reasons (I consider myself to be a person who always has strong beliefs, etc.) but we are wrong about this, and that is what this test have shown. For some of the questions, the different formulations of the questions made us interpret them slightly differently, and this interpretation affected our statement.

Now, a result where we fail this test even more miserably would of course indicate that the inner sense theory was incorrect. The problem is that the competing theory-theory is not much better equipped at explaining complete failures at such a test. While it can account for that we

[8] It might be that it is very hard to construct two different formulations of the same question, without the test subjects reading it as two different questions, connected to different beliefs. This might also have been the case in the "choice blindness"-tests. I will ignore both these possibilities.

sometimes make errors, the theory is not compatible with that we *usually* make errors. If we fail this test, neither theory have a good explanation. This result would indicate that beliefs are not at all as stable as we think they are. If this is the case, the inner sense theory could still be true, and the unstable nature of beliefs would be the reason for the complete failure. What is more, I do not think we would fail this kind of test, I think we would do very well, and that would not give us any justification for either theory. That, however, is of course an empirical question.

These examples of possible tests are not exhaustive, of course. I presume that several more testable implications would appear if these competing theories were spelled out in much more detail. What could be clarified is, for instance: a/when and why the inner sense would fail, b/the nature of belief, so that the statements could be about an actual belief and avoid transformation-related problems, and c/how the inner sense cooperates with the "reasoning" or the "backup-method".

As a final note, I might add that in most of the debate between proponents of the inner sense-theory and the theory-theory, the focus is not on choice blindness-tests; instead the debate has mostly concerned conditions, such as autism and schizophrenia, and cognitive development of children. Evidence from these research fields might provide sufficient reasons to believe in either theory. However, the experiments described here are special in that they are designed to test these kinds of theories. More focus has been on the explanative power of these theories, rather than the predictive power. I believe a shift to a discussion about testing the predictions of the theories would benefit the debate.

5. Conclusion

This essay argues that the choice blindness experiments provides evidence that an inner sense that we use often, automatically and without any particular effort, is implausible. If the inner sense theories suppose that the inner sense has such characteristics, we should not fail the choice blindness-test, and since we do we should favour a different theory. However, an inner sense theory

is not committed to this view. First and foremost, it can claim that while the inner sense provides privileged access to our beliefs, some beliefs are harder to introspect than others, meaning that it takes more time and effort to form second-order beliefs about them. This might be compared to how some memories are harder to retrieve, even though we have direct and non-inferential access to them. Second, the inner sense theory can state that we are reluctant to use our inner sense to introspect beliefs that are hard to introspect. Third, when we are reluctant we need a good reason to use our inner sense, such as suspicion or uncertainty, and we are given no such reason. Arguing for your belief is not in itself a secure way to make the inner sense work when it comes to introspecting "hard" beliefs, and the discussion reviewing phase does not raise any suspicions.

If this is the case, then we could be choice-blind in certain situations despite the existence of an inner sense. One such situation is when we are deceived. The fact that we can be deceived does not show that we do not use an inner sense. It merely shows that the inner sense is not infallible. One reason why it is fallible is the "cognitive economy-argument". However, if we are to favour the inner sense theory, we would need to do more than just show that it is able to avoid the problematic implications of empirical tests. The theory-theory is in this sense stronger, since it predicts the result given in the choice blindness tests. We would like the inner sense theory to give new predictions that we could test. It is hard, however, to come up with a test that we would pass if and only if the inner sense theory was true, or for that matter that we would fail if and only if it was false. However, it might be possible, and I believe it to be a fruitful line of inquiry.

6. References

Armstrong, D. M., 1971. *A Materialist Theory of the Mind.* 3rd ed. London: Redwood Press Limited.

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

Dretske, F., 2004. Change Blindness. *Philosophical Studies,* Volume 120, pp. 1-18.

Gertler, B., 2011. *Self-Knowledge.* London: Routledge.

Gertler, B., 2014. *Self-Knowledge.* [Online]
Available at: http://plato.stanford.edu/archives/win2014/entries/self-knowledge/
[Accessed 18 September 2015].

Goldman, A. I., 2006. *Simulating Minds: The Philsophy, Psychology, and Neuroscience of Mindreading [Electronic resource]*. New York: Oxford University Press.

Gopnik, A., 1993. How we known our own minds: The Illusion of first person knowledge of intentionality. *Behavioral and Brain Sciences,* Volume 16, pp. 1-14.

Hall, L., Johansson, P. & Strandberg, T., 2012. Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoSONE,* 19 September.7(9).

Hall, L., Johansson, P., Tärning, B. & Sikström, S. D. T., 2010. Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition,* 117(1), pp. 54-61.

Hall, L. et al., 2013. How the Polls Can Be Both Spot On and Dead Wrong: Using Choice Blindness to Shift Political Attitudes and Voter Intentions. *PLoSONE,* 10 April.8(4).

Johansson, P., 2006. *Choice blindness [Electronic resource] : the incongruence of intention, action and introspection.* Lund: Diss. Lunds universitet.

Johansson, P., Hall, L., Sikström, S. & Olsson, A., 2005. Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task. *Science*, 7 October, 310(5745), pp. 116-119.

Nichols, S. & Stich, S. P., 2003. *Mindreading [Elektronisk resurs] An Integrated Account of Pretence, Self Awareness, and Understanding Other Minds.* Oxford: Oxford Univesity Press.

Nisbett, R. E. & Wilson, T. D., 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review,* May, Volume 84, pp. 231-259.

Pärnamets, P. H. L. S. T. B. C. &. J. P., submitted. Looking at choice blindness: Evidence from eye-tracking and pupil dilation.

Schwitzgebel, E., 2011. Knowing Your Own Beliefs. *Canadian Journal of Philosophy,* Volume 35, pp. 41-62.

Schwitzgebel, E., 2014. *Belief.* [Online]
Available at: http://plato.stanford.edu/archives/spr2014/entries/belief/
[Accessed 16 03 2015].

Simon, D. J. & Chabris, C. F., 1999. Gorillas in Our Midst: Sustained Inattentional Blindness for Dynamic Events. *Perception* , Volume 28, p. 1059–1074.

Stich, S. P. & Nichols, S., 1998. Theory theory to the Max. *Mind and Language,* 13(3), pp. 421-449.

Tversky, A. & Kahneman, D., 1974. Judgment under Uncertainty: Heuristics and Biases. *Science,* 27 September, 185(4157), pp. 1124-1131.