Choice Blindness: The Incongruence of Intention, Action and Introspection

Johansson, Petter

2006

Choice Blindness

Choice Blindness

The incongruence of intention, action and introspection

Petter Johansson

Lund University Cognitive Studies 130

For further information, see authors' webpage:
www.lucs.lu.se/people/petter.johansson

Cover and book design: David de Léon
Cover photo: Mark Hanlon

Printed in Sweden, Media-Tryck
ISBN 91-628-6954-X
ISSN 1101-8453
ISRN LUHFDA/HFKO-1017-SE
Lund University Cognitive Studies 130

To my family, and my family of friends

"Era uma vez, em um reino distante, cientistas mostraram a voluntários alguns pares de fotografias de rostos de mulheres. 'Qual lhes parece mais atraente?' os cientistas perguntavam. Quando o voluntário revelava sua escolha, os cientistas então pediam que ele descrevesse verbalmente as razões para explicar sua escolha.
Mas o que os voluntários não sabiam é que os cientistas eram traquinas, e algumas vezes usavam um passe de mágica para trocar as fotos depois que a escolha havia sido feita. Assim, pediam ao voluntário para explicar por que havia escolhido o rosto que, em verdade, não havia escolhido."

ꕥ

"It was a time, in a distant kingdom, scientists had shown to the volunteers some pairs of photographs of faces of women. 'Which them seems more attractive', the scientists asked. When the volunteer disclosed its choice, the scientists then asked for that it described the reasons verbally to explain its choice.

But what the volunteers did not know it is that the scientists were traquinas [rascals], and some times used a magician pass to change the photos later that the choice had been made. Thus, they asked for to the volunteer to explain why it had chosen the face that, in truth, it had not chosen."

An article about choice blindness in Portuguese, automatically translated to English through Babelfish

CONTENTS

Publication Histories

The publication histories for the papers included in the thesis are as follows:

Paper one

Paper one is based on the following presentations:

Johansson, P., Hall, L., & Olsson, A. (2004). *From change blindness to choice blindness*. Towards a Science of Consciousness 2004, Tucson, Arizona, April 7–11, 2004.

Hall, L., Johansson, P., Olsson, A., & Sikström, S. (2004). *Choice blindness and verbal report*. The Association for the Scientific Study of Consciousness, 8th Annual Meeting, University of Antwerp, Belgium, June 25–28, 2004.

Johansson, P., Hall, L., Olsson, A., & Sikström, S. (2004). *Facing changes: choice blindness and facial attractiveness*. The 28th International Congress of Psychology, Beijing, China, August 8–13, 2004.

Paper two

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*, 116-119.

Paper three

Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (in press). How something can be said about telling more than we can know. *Consciousness and Cognition.*

Paper four

Hall, L., Johansson, P., Tärning, B. Deutgen, T., & Sikström, S. (2006). Magic at the marketplace. *Lund University Cognitive Studies, 129.*

Acknowledgements

I have during the years used many different images to describe what it is like to write a thesis. My present favourite is from Werner Herzog's movie *Aguirre: The Wrath of God*, starring Klaus Kinski. Against all advice, he sets out on a small raft to find El Dorado – the legendary land of gold. The only map used is his hunches and hopes. Carried by the currents of the Amazon River, he travels further and further into the wilderness. But the quest is a disaster. There are no signs of gold; there are no signs of anything. Each day the same, day after day. The river runs fast, but only in one direction. And there is no turning back. In the end, he stands alone on the raft. Stark mad, raving about future riches and rewards.

I now see that my case differs in at least two respects. I may not have found gold, but at least I came ashore. The journey ends here. Secondly, I have not been alone on the raft. Rather than dying off, the crew has been constantly growing. So the following are the people I would like to thank for cheering me on or keeping me company on this trip.

First of all, Peter Gärdenfors, professor of the department of Cognitive Science at Lund University and my first supervisor. I would actually like to thank him most for what he did last. Our long discussions regarding the rhetorical composition of the introduction and the thesis were both fun and very valuable. On a more general level, I would like to thank him for creating the special atmosphere of enthusiasm and intellectual curiosity we have at the department. This is something he is often praised for, but it is a credit he deserves.

Next in line is my current supervisor Sverker Sikström. I would like to thank him for his empirical know-how and the scientific rigor he added to our team. You might not have noticed, but I have learnt a lot of things from you.

I would also like to thank all the past and present members of the Cognitive Science department, for all the seminars and the discussions we have had over the years: Martin Bergling, Petra Björne, Nils Dahlbäck, Philip Diderichsen, Pierre Gander, Agneta Gulz, Kenneth Holmqvist, Jana Holsánová, Nils Hulth, Birger Johansson, Magnus Johansson, Paulina Lindström, Petter Kallioinen, Peter Kitzing, Lars Kopp, Maria Larsson, Jan Morén, Mathias Osvath, Tomas Persson, Annika Wallin and Jordan Zlatev.

Christian Balkenius deserves special credit for helping us with the diagrams for Paper 2 in the thesis, a template I have shamelessly used for all work done since. He has also been my interface to the world of Mac, and has given priceless help and advice when I have been in a tight spot the hours before a deadline without a clue what to do.

I would also like to thank our brilliant secretary (well, you are the best!) Eva Sjöstrand, for always taking care of everything that needs to be taken care of.

There are also a number of people I have collaborated with when writing the papers in this thesis. Andreas Olsson is co-author on Paper 2, and I thank him for his input on how to best go from our results and ideas to a proper article. Betty Tärning, co-author on paper 3 and 4, has been invaluable as she has performed experiments, catalogued data, as well as transcribed verbal reports. Apart from this, she has contributed greatly in our often long and sometimes seemingly aimless group discussion on what to do with choice blindness. Andreas Lind was essential in the making of Paper 3, both with his linguistic expertise as well as his tireless devotion to getting things done (I agree, it is more fun to work hard). Finally, Thérèse Deutgen played a large part when making magic at the marketplace for Paper 4.

I would also like to thank a number of people at the department of linguistics at Lund University for all their help and advice concerning the linguistic analyses performed in Paper 3: Mats

Andrén, Victoria Johansson, Joost van de Weijer and Jordan Zlatev. While some of you might not agree with our interpretation of the data, I hope you still thought our project was of some interest.

In addition to those that had to, several people read and commented on my introduction. I would like to thank you all for your time, I really appreciated your input: Lars Brink, Ingegärd Johansson, Markus Karlsson, Peter Kitzing and Björn Peterson.

The people that have meant most to me in my daily life writing this thesis are David de Léon, Jens Månsson and Lars Hall. Through the long lunches and late nights, you guys have made being at work fun. I especially thank David for all the artefactual intelligence you put into the making of this book. I admire your eye for beauty, and share your appreciation for the good in life. And Lars, I am not going to even try to list the things I thank you for. You are a brilliant man, and my dearest friend.

*

There are also several people and institutions I would like to mention outside the immediate environment of Lund University.

I would like to thank *Bank of Sweden Tercentenary Foundation* for funding large parts of my Ph.D. studies. This was a separate project from the thesis, which resulted in the edited volume Gärdenfors and Johansson (2005). *Cognition, Education and Communication Technology*. Lawrence Erlbaum Associates.

I would similarly like to thank *The New Society of Letters at Lund* for financing a year at University of East London.

I thank UEL for hosting me that year, and especially my personal host Tom Dickins for making my stay very rewarding on a both personal and professional level. He taught me that "Evolution is the Way and the Truth" and that "Clarity is All" – I hope he is not too disappointed by the near absence of both these things in the thesis. At EUL I also realised that the pub is the best place for academic discourse, and in addition to Tom I would also like to especially thank Eike Adams, Chris Pawson and Qazi Rahman for teaching me this.

Professor Nick Humphrey at London School of Economics invited me to attend his weekly work-in-progress seminar on evolutionary psychology. I thank him for this as well as for letting me present my own ideas at this forum.

I would also like to thank Julie McNiff, for giving me warmth and shelter, and for showing me how London should be lived.

*

Peter Rosengren guided us through the fine arts of card magic; I thank him for this (and for not laughing at our clumsy efforts). Having a real magician in the group has been the source of envy for colleagues all over the world. I would also like to thank Karl Berseus and Axel Adlercreutz for our long discussions at the "Swedish Magicians' Circle" conference, regarding how to make objects disappear.

Anders Hall deserves special credit for all his help with my computers over the years. The final act was his efforts the last day before handing in my thesis (earlier this morning!), when Windows suddenly decided to implode. He saved me today, as he has done many times before.

I would like to thank my old friend Magnus Dittmer for all the summer weeks I spent with you and Marie, and for giving me some perspective on life – there are many ways to live it happily.

The last few years, I have shared apartment with Markus Karlsson, Hans Appelqvist, Hilderun Gorpe and Malin Skoglund. I would like to thank them for being part of a home that has made me want to return from work as well.

*

Finally, I would like to thank my closest family.

First of all, my girlfriend Marie. To be we is the best part of my life. Finishing this thesis was always going to be hard, but knowing that I had already won made it so much easier. You know I will always point at your picture.

My aunt Eva and my grandmother Elsa, for all the light and laughter, and for caring so much for me.

My sister Lovisa and her family: Lars, Rasmus and Klara. For letting me share your space, the warmest and safest place in the world.

My father Leif, who taught me the beauty of obsession. It is not what you do but how you do it that matters.

My mother Ingegärd and Bo, who taught me to love thoughts, but also convinced me to try to put some data in. You are right: no one will listen if I just talk. Thank you mum, for always being there.

And to not exclude anyone, I would also like thank myself, for pulling this off without falling apart.

Introduction

Look at the two faces on the book cover. Try to decide which one of them you find more attractive. After you have made up your mind, focus on the face you preferred, and explain to yourself why you liked that one better. Now imagine I told you that you actually preferred the other face. After your decision – but before you started to talk – I switched the position of the pictures, so you are now looking at the face you did *not* choose. When you gave your reasons you were in fact looking at the opposite of your choice.

Would you take my word for it, or would you find it hard to believe?

If you think you would have noticed the manipulation you are not alone; this is what most people think. But however unlikely it may seem, it is not at all certain that you *would* have seen the switch. And had you not seen when the pictures changed places, you are also quite likely to give a long and elaborate description of why you chose this face and not the other.

Despite its brevity, this scenario contains all the major components of the thesis. The work presented is an empirical and theoretical exploration of the finding that people are prone to miss even dramatic mismatches between what they want and what they get. The fact that it is possible to manipulate the relation between people's intentions and the outcome of their actions *without them noticing* is what my collaborators and I have dubbed *choice blind-*

ness. This effect is demonstrated in a series of experiments, using both different stimuli and different experimental methods.

But not only were the participants in our experiments blind to the manipulation of their choices, they also offered introspectively derived reasons for preferring the alternative they were given instead. The second major component of this thesis is thus the participants' verbal reports explaining choices they did not intend to make. These reports are analysed both in isolation and in relation to reports from non-manipulated choices. By comparing the content of the verbal reports with the properties of the chosen items it is possible to establish that the reports are sometimes "confabulatory" – i.e. when the participants refer to unique features of the initially non-preferred face (e.g. a pair of earrings) as being the reason for choosing this alternative rather than the other. As an additional finding, the reports stemming from manipulated choices seem to be just as rich and elaborate as the ones given in non-manipulated trials.

Finally, I consider the experimental methodology to be a finding of its own. We have created a number of different experimental procedures in which we generate a mismatch between what the participants intend to choose and the outcome they experience as being their choice. By using a binary choice task, we can always be certain that our participants actually wanted the opposite of what they were given. All the empirical work presented shares these general characteristics.

Thus, the three things I see as novel in the thesis are the choice blindness effect, the verbal reports based on manipulated choices, and the experimental approach as such. Throughout the book and the rest of the introduction, this is what it is all about.

In this introduction, each of the four papers is presented with a very compressed descriptive recapitulation of the experiments, the results and the conclusions drawn. The papers are then discussed in terms of related topics and theory, organised around the three major themes identified above.

I consider Paper 1 and 2 as well as the *Supporting Online Material* accompanying Paper 2 as belonging to the same project, and they will be summarised and discussed together in relation

to the choice blindness effect. The theoretical backdrop for this discussion is the nature of *Folk Psychology*, and the use of belief-desire explanations in cognitive science modelling. I will argue that our results represent a substantial problem for philosophers and cognitive scientists that connect their models too closely to a Folk Psychological model of the mind. As such, the choice blindness effect challenges the commonsense assumption that beliefs, desires and intentions, are entities in the brain. Instead, our results are better understood within the framework of the *Intentional Stance* (Dennett, 1987), in which beliefs and desires are seen as predictive tools we use in our attempts to make sense of ourselves and others.

The discussion of Paper 3 will be focused on the analyses of the verbal reports. The natural context of this discussion is the perennial battle in psychology and philosophy regarding the validity of introspective self-reports. Extra attention is given to the debate following the publication of Nisbett and Wilson (1977), an article which strongly questions the accuracy of introspection. I will argue that while our results can be given a similar interpretation as was given Nisbett and Wilson's, our experimental method is a significant step forward. Still, one conclusion must be that that our results indicate that we know a lot less about ourselves than we think we do.

In relation to Paper 4, I expand on the idea of using our experimental approach as a more general research tool, and give a glimpse of future studies planned.

Choice Blindness

I consider Paper 2 to be the centrepiece of the thesis, and the paper best served to introduce the approach as a whole. I will therefore start with the summary of Paper 2.

Summary Paper 2: ***Failure to detect mismatches between intention and outcome in a simple decision task.*** The participants in the study of Paper 2 were shown two pictures of female faces, and were instructed to point at the face they found most attractive. After pointing, the chosen picture was given to the participants, and they were asked to explain why they preferred the picture they now held in their hand. Unknown to the participants, using a double-card ploy, the pictures were sometimes covertly exchanged mid-trial. Thus, on these trials, the outcome of the choice became the opposite of that intended by the participants (see Figure 1 in Paper 2).

Each of the 120 participants performed 15 choice trials, of which three were manipulated. The time given to make a choice, and the similarity of the face-pairs were varied. For time, three choice conditions were included: one with two seconds of deliberation time, one with five, and a final condition where participants could take as much time as they liked. For similarity, a high and a low similarity set of target faces was used.

A trial was classified as detected if participants showed any signs of detection in immediate relation to the switch (such as explicitly reporting that the faces had been switched, or indicating that something went wrong with their choice), or if the participants voiced any suspicion in the debriefing session after the experiment.

Counting all forms of detection across all experimental conditions, no more than 26% of the manipulated trials were detected. There were no significant differences in detection rate between the two groups of stimuli used. For viewing time, the 2-second and 5-second conditions did not differ in detection rate, but there were significantly more detections in the free viewing time condition.

The verbal reports were also recorded, transcribed and analysed. Of primary interest is the relation between the reports given in manipulated and non-manipulated trials. In the non-manipulated trials the participants just answered why they had preferred the chosen picture, but when doing the same thing in the non-detected manipulated trials the participants described and gave reasons for a choice they did not intend to make. The two classes of reports were analysed on a number of different dimensions, such as the level of emotionality, specificity and certainty expressed, but no substantial differences between manipulated and non-manipulated reports were found.

The experiment also established the extent to which a report could be matched to the picture originally chosen or to the manipulated outcome received – i.e. if the participants talked about

the face they thought more attractive first or the one they ended up with after the switch was performed. The conclusion drawn in Paper 2 is that the relationship between intention and outcome may sometimes be far looser than current theorising has suggested. As such, choice blindness warns of the dangers of aligning the technical concept of intention too closely with commonsense. The analyses of the verbal reports shows that in some trials we can be certain that the participants confabulate or construct their answers in line with the manipulations made, as they refer to unique properties of the initially non-preferred face. The lack of differentiation between the manipulated and non-manipulated reports casts doubt on the origin of the non-manipulated reports as well; confabulation could be seen to be the norm and truthful reporting something that needs to be argued for.

The Supporting Online Material functions as an appendix for Paper 2. Several aspects are expanded and detailed, such as the experimental procedure, statistical measures used, detection criteria, the analyses of the verbal reports, and the relation to previous studies.

Summary Paper 1: ***From change blindness to choice blindness.*** Paper 1 is a precursor to Paper 2, in terms of both theory and empirical method. The participants either had to choose which of two abstract patterns they found most aesthetically appealing or which of two pictures of female faces they found most attractive. Fifteen trials were used, of which three were manipulated. The choice task was presented on a computer screen, and the participants had to indicate their choice by moving the cursor to the chosen picture. When all the choice trials were completed, an unannounced memory test was introduced. The participants had to look at all the pairs again, without time-constraint, and try to remember which face or pattern they previously preferred. The result was similar to that in Paper 2, as the participants showed considerable levels of choice blindness. The memory test revealed that the participants had been influenced by the manipulations made, and tended to remember the manipulated outcome as the alternative they originally preferred.

Surprise, surprise

So why do I think our experimental results are an interesting finding? From a commonsense perspective, choice blindness seems a baffling phenomenon. How can someone choose *x*, and then not notice when given *y* instead? Do we not know what we want when we make a choice? Given the lack of similarity between the faces (see Picture 1 in SOM), how is it possible *not* to notice if they are swapped? This does not seem to fit well with our ordinary intuitions of how we function.

But it is not just the description of the experiment and the results that people find surprising. In the debriefing session after the experiment in Paper 2, all participants were asked a series of increasingly specific questions to investigate whether they suspected in any way that something had gone wrong ("What did you think about the experiment?", "Did you find anything odd about the experiment?" and "Did you notice anything strange about the stimuli presented in the experiment?"). Participants who revealed no signs of detection were then presented with a hypothetical scenario describing an experiment in which the faces they choose between are secretly switched (i.e. the very experiment they had just participated in), and asked whether they thought they would have noticed such a change. The result shows that, of the participants who failed to notice any of the manipulations, 84% believed that they *would* have been able to do so. Accordingly, many participants also showed considerable surprise, even disbelief at times, when we debriefed them about the true nature of the design. We call this effect "choice blindness blindness"; i.e. the overconfidence in our own ability to detect choice-manipulations (For a similar meta-cognitive error in relation to change blindness, see Scholl, Simons, & Levin, 2004). In my opinion, this is also the strongest evidence there is that we have discovered something genuinely contra-intuitive.[1]

Our commonsense intuitions are also a good starting point for a more theoretically grounded discussion of choice blindness. In philosophy and cognitive science, the totality of our everyday psychological explanations is referred to as Folk Psychology (Bogdan, 1991; Christensen & Turner, 1993; Greenwood, 1991). When we try to make sense of other people, or when we answer

1. This is also a strong argument with regards to the question how we can know that the participants really did not detect the manipulations. Maybe the participants saw all manipulations but just did not tell us? But to first confidently claim that they think they would have noticed a switch, and then "feign" surprise and deliberately lie when asked if they saw the manipulations, is something that seems a very odd thing to do. In addition, counting all experiments mentioned or described in this thesis, we have tested around 470 participants and classified around 790 trials as non-detected choice manipulations. It does not seem likely that we have misclassified all of them. The issue of forms and levels of detection is further discussed in Supporting Online Material and in an interchange with a commentary on Paper 3 (Hall, Johansson, Sikström, Tärning & Lind, in press; More & Haggard, in press).

questions such as *why* we preferred one picture over another, we phrase these answers in mental state descriptions such as beliefs and desires. For example:

"She must think that no one can see her through the window."
"Probably, he just really really wanted that apple."
"I thought you thought that I believed you to be innocent."

We are all experts on Folk Psychology, it is a language we are fluent in from a very early age. When examined more closely, Folk Psychological descriptions have certain law-like regularities, such as:

If X wants Y and believes that it is necessary to do Z to get Y, X will do Z

If Petter wants ice-cream and believes it is in the freezer, he will open the freezer and take one out. But they work as explanations as well as predictions – if Petter is seen opening the freezer and taking an ice-cream, he most likely wants ice-cream as well. We use the framework of Folk Psychology all the time, to understand and make sense of both ourselves and others. We believe, desire, intend, want, hope, think, fear, etc. But despite being a seemingly indispensable tool for understanding and interacting with each other, our Folk Psychological constructs are problematic entities. What exactly *are* beliefs, desires and intentions?[2]

2. But we sometimes feel the limits of Folk Psychology. In the 110th minute of the 2006 world cup final, Zinedine Zidane suddenly head-butts Marco Materazzi and is sent off. The most celebrated player of the modern era; the captain of the French team; a true hero of the people. He declared that he would retire after the tournament, and then defied age and expectations and played some of the best games of his career. And in the last act, he puts his entire legacy at risk. More than a billion spectators sit stupefied in front of the television set. *Why did he do it?* In the replay it is clear that the Italian defender says something. Zidane hesitates for almost a second, as if contemplating the alternatives, and then charges. The only possible explanation is the words said, but how can they have had the force to make him do what he did? From a Folk Psychological perspective, it is interesting to note that everyone agreed that it *must* have been something extremely offensive or vile, or some deeply personal matter. The magnitude of the insult does not only need to match the future consequences disregarded, it is also the personality we have pinned on Zidane after getting to know him for 15 years watching him play. Still, in this case, it does not feel like we will ever understand the action.

The Great Divide

The status and nature of Folk Psychology is an old battleground in philosophy of mind and cognitive science, crisscrossed with trenches and fronts opened in all directions. What everyone seems to agree on is that Folk Psychology is a very powerful tool in explaining and predicting people's behaviour, but apart from that, they disagree on just about everything else. Philosophers argue about how well it actually functions as a coherent scientific theory (Churchland, 1981), while developmental psychologists disagree on both how and when we acquire the mental concepts we use in later life (Astington, 1993; Gopnik, 1993). Primatologists discuss to what extent our near neighbours share our belief-desire type of "theory of mind" (Premack & Woodruff, 1978), while others argue whether a "theory of mind" is a prerequisite for the development of a Folk Psychology in the first place (Baron-Cohen, 1994, 1995).

But there are two main threads in the debate. What do we *do* when we understand each other using the conceptual framework of Folk Psychology, and to what extent does the theory of Folk Psychology correspond to what is actually going on in the mind (or the brain)? Regarding the first question, there are two major positions: You are either a theory-theorist and argue that we apply the *theory* of Folk Psychology as any other theory when we explain what people do (Gopnik, 1993), or you are a simulation-theorist and argue that we primarily understand each other through a kind of mental role-playing in which we put ourselves in other people's position and thereby "experience" what mental states they are likely to have (see e.g. Goldman, 1993). I will return to this question briefly when discussing introspection in the next chapter summary, but in relation to choice blindness the second question is more important. So, in what sense do the entities of Folk Psychology such as beliefs, desires and intentions exist?

The philosophical position that assumes Folk Psychology to describe real things residing in the head is called Intentional Realism, and the foremost champion of this doctrine is Jerry Fodor (1983, 2000). According to him, it is Folk Psychology all the way in. The reason Folk Psychology works so well is because it happens to be true. In a distant future, when we have mapped out the workings of the brain, we will find the equivalents of beliefs and desires.

They will be discovered to be the fundamental building blocks in the internal cognitive machinery that governs our behaviour. He is an adamant defender of this position, and he does not take his job lightly: "if commonsense psychology were to collapse, that would be, beyond comparison, the greatest intellectual catastrophe in the history of our species" (1987, p. xii). Despite this, I have my allegiances elsewhere.

Daniel Dennett explains both what Folk Psychology is and how we use it within the same theoretical framework: The Intentional Stance. Some additional background is necessary to appreciate his position. Dennett (1987) presents a taxonomy of stances or viewpoints from which to predict or understand any system. First we have the Physical Stance, from which systems are predicted by exploiting information about their physical constitution. Since, in the end, humans are nothing more than extremely complex physical systems we are in principle predictable with this method. Next we have the Design Stance, from which one understands the behaviour of a system by assuming it is composed of elements with functions, i.e. that it has a certain design, and that it will behave as it is designed to do under various circumstances. Finally, there is the Intentional Stance, from which one predicts a system by treating it as an approximation of a rational agent. We attribute the beliefs, desires and goals it ought to have, taking into consideration previous actions, verbal statements and available options[3].

3. To complicate things further: in philosophy there is also an underlying debate on the nature of *intentionality*, which is a technical concept referring to the ability of one thing to be *about* something else. The word "turnip" refers to a specific vegetable; a turnip as such can not refer. Apart from words and symbols, mental entities can also be about other things: I believe there is gold at the end of the rainbow, I think about what to eat for lunch. Brentano (1874/1973) famously stated that as physical objects cannot be about other things, but mental states can, mental states cannot be reduced to physical states or entities – *the irreducibility of the mental*. This can either be interpreted as supporting some form of dualism (Chisholm, 1966); or that in an absolute sense, mental states do not exist and therefore we cannot have a proper science about them (Quine, 1960). Both Fodor and Dennett opt for other alternatives: Fodor claims that mental states *are* physical states and get their meaning or content through causal links to the objects they refer to, while Dennett agrees with Quine that beliefs and desires do not exist as objects, but claims them to exist as relations seen from the intentional stance, whose content ultimately can be derived from the rationality presupposed by an evolutionary perspective. Much abbreviated – depending on your perspective this debate is either extremely important or largely irrelevant for the present thesis, but I will nevertheless relate it no further here.

The Intentional Stance is in Dennett's view the backbone of our Folk Psychology, and it is the rationality assumption that is the guiding principle when we create a psychological explanation. In his own words:

> Here is how it works: first you decide to treat the object whose behavior is to be predicted as a rational agent; then you figure out what beliefs that agent ought to have, given its place in the world and its purpose. Then you figure out what desires it ought to have, on the same considerations, and finally you predict that this rational agent will act to further its goals in the light of its beliefs. A little practical reasoning from the chosen set of beliefs and desires will in most instances yield a decision about what the agent ought to do; that is what you predict the agent will do. (Dennett, 1987; p. 17)

Within this framework, every system that can be profitably treated as an intentional system by the ascription of beliefs, desires, etc., also *is* an intentional system in the fullest sense (see Dennett, 1987; 1991a). Not just human beings but countries, banks, butterflies – even the lowly thermostat – have beliefs and desires if we gain any predictive leverage from ascribing such states to them. Dennett is thus very inclusive regarding what can be considered to *have* beliefs and desires, as well as what should be considered to *be* beliefs and desires. They exist as patterns in the world, to be seen from the Intentional Stance (Dennett, 1991b). With this perspective, it is not surprising that he does not think that belief-desire prediction reveals the exact internal machinery responsible for the behaviour.

> We would be unwise to model our scientific psychology too closely on these putative *illata* (concrete entities) of folk theory. We postulate all these apparent activities and mental processes in order to make sense of the behavior we observe – in order, in fact, to make as much sense possible of the behavior, especially when the behavior we observe is our own […] each of us is in most regards a sort of inveterate auto-psychologist, effortlessly *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt) good theorizing. (Dennett, 1987; p. 91, emphasis in original)

When we explain our own behaviour in terms of Folk Psychology, we do this by applying the Intentional Stance towards ourselves as well. I observe myself and interpret my actions rather than getting to know my beliefs and desires from the inside. And after explaining a certain act and having clad my behaviour in words, the description of the mental entities deemed responsible for my ac-

tions now has a concrete existence not previously enjoyed: "The intentions are as much an effect of the process as a cause – they emerge as a product, and once they emerge, they are available as standards against which to measure *further* implementation of the intentions" (Dennett, 1991a, p. 241, emphasis in original).

If we take a look at our experiment, the behaviour of the participants seems to make sense given Intentional Stance theory. What the participants seems to be doing is to make interpretations. They see themselves act, and assume that the picture they reached for and were given also was the picture they intended to choose. All the external evidence points in this direction, it is a reasonable conclusion to draw given the circumstances. They took the card, so they must have wanted it. But the things they say do not need to be actual descriptions of what went on in their heads prior to the decision. Some kind of decision-making process made them choose one face over the other, but the "reasons" responsible for this do not need to correspond to the things they say. And there need not be any higher-order intention *in the brain* to choose one face over the other, the outcome of the internal evaluation might only result in the motor act of pointing to the face preferred. The reasons the participants give is their own interpretation of *why* they must have wanted this picture rather than the other. In a sense, they inform themselves as much as everybody else about what they wanted when they perform and then explain their actions.

It is of course hard to draw any strong ontological conclusions from our experiments; it would be silly to say that we have shown Intentional Realism to be false. But it is also quite evident that our results better fit Dennett's perspective than Fodor's. If Folk Psychology is an instrument of interpretation, it should be possible to make "mistakes" about ourselves – e.g. to make a belief-desire interpretation that does not fit with the lower-level implementation of the action. As it now stands, one possible explanation why the participants in our experiments did not detect the mismatch between their intentions and the outcome of their actions could simply be that the prior intentions (as conceptualised by Intentional Realism) do not exist. Intentions are not well specified concrete entities; they are abstractions we use to make sense of

behaviour. There are processes in the brain that are responsible for the evaluation that led to the action, but there is no well-specified internal description of what the participants intended to do in addition to that. Something must precede the action, but that process does not need to exist in a format that is comparable to the Folk Psychological description of what went on.

But in relation to our experiments, the problem for Intentional Realism becomes more vivid when we leave the high grounds of philosophical controversy and instead look at more specific cognitive science models of human behaviour. Even if not explicitly endorsed, Intentional Realism about Folk Psychological constructs is a ubiquitous feature in cognitive science. In line with the reasoning of Fodor, many researchers have taken the apparent success of Folk Psychology as evidence that there must be corresponding processes in the brain that closely resemble the goals and intentions postulated by the theory.

Letting the intentions out of the box

In cognitive psychology and cognitive science, a frequently used tool for describing cognitive and behavioural relations is the flowchart model. When it comes to goal-directed behaviour, one thing the models often have in common is that in the uppermost region of the chart, a big box sits perched governing the flow of action. It is the box containing the Prior Intentions (Brown & Pluck, 2000; Jeannerod, 2003). In these models, intentions and goals are discrete entities with very specific identifiable properties.

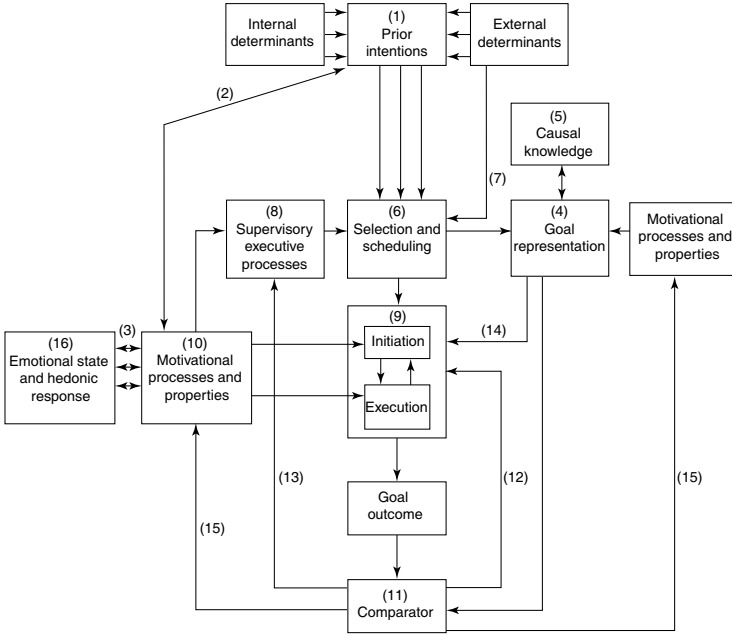

Figure 1. Model of goal-directed behaviour, from Brown and Pluck (2000).

The model above is labelled a neuro-cognitive and a neuro-philosophical formulation of goal-directed behaviour (Brown & Pluck, 2000; Jeannerod, 2003). The model is in itself a synthesis of other models from several different areas in cognitive science, such as cognitive and functional anatomy of will and volition (Ingvar, 1999; Spence & Frith, 1999), neurobiology of reward (Schultz, 1999), and philosophical descriptions of purposeful behaviour (Searle, 1983).

According to this flow-chart, a goal-directed action is driven by the Prior Intention. For an action to be goal-directed the system needs to have an internal representation of the goal, as well as knowledge of particular actions that will lead to achieving the goal. The action is controlled through feedback from the Comparator, which compares and evaluates the goal outcome against the goal representation. The output from the Comparator

is used to maintain or stop the ongoing action, and will further influence the motivational processes involved in the task.

It is in relation to a model like this that choice blindness as a phenomenon becomes very hard to account for. Arguably, choosing and taking the more attractive of two pictures of faces must be considered a goal-directed behaviour. The action performed in our experiment has all the components of the model, but still the mismatch between the intention and the outcome is not detected. The Comparator should have stopped the process when the participants received the opposite of their choice, but it didn't. How can this be?

One explanation could be that in models like this, the internal representation of the goal state only concerns low-level features; maybe they are only meant to describe actions on a motoric level, such as reaching for the remote or tying one's shoes. But Brown and Pluck (2000) do not put any restrictions on the kinds of actions or level of goal specificity that this model is supposed to handle:

> Within neuroscience, the construct of GDB [goal-directed behaviour] is increasingly being used to operationalize a broad spectrum of purposeful actions and their determinants, from the simplest single-joint movement, to the most complex patterns of behaviour. GDB is construed as a set of related processes by which an internal state is translated, through action, into the attainment of a goal. The 'goal' object can be immediate and physical, such as relieving thirst, or long-term and abstract, such as being successful in one's job or the pursuit of happiness. (p. 416)

Apparently, both intentions and goals can be both abstract and complex, and for the Comparator to fill any function it must be able to detect when the higher-order goals are obtained or not.

Another possible objection to protect the model is that the Comparator just did not do its job this time. Maybe checking the relation between intentions, goals and results is optional rather than essential? But the ability to compare the prior goals with the outcome obtained is an ever-present feature in action modelling, and is seen to be fundamental for a great number of things:

> Our ability to judge the consequences of our actions is central to rational decision making […] A key component to survival in a constantly changing environment is the ability to evaluate the consequences of one's actions and to adapt one's behavior accordingly. (Walton, Devlin, & Rushworth, 2004; p. 1259)

> Flexible behavior requires a system for relating responses to the current context and one's goals. (Badre & Wagner, 2004; p. 473)

> Adaptive goal-directed behavior involves monitoring of ongoing actions and performance outcomes, and subsequent adjustments of behavior and learning. (Ridderinkhof, Ullsberger, Crone, & Nieuwenhuis, 2004; p. 443)

> [T]he anterior cingulate cortex (ACC) has a fundamental role in relating actions to their consequences, both positive reinforcement outcomes and errors, and in guiding decisions about which actions are worth making. (Rushworth, Walton, Kennerley, & Bannerman, 2004; p. 410)

> Flexible adjustments of behavior and reward-based association learning require the continuous assessment of ongoing actions and the outcomes of these actions. The ability to monitor and compare actual performance with internal goals and standards is critical for optimizing behavior. (Ridderinkhof, van den Wildenberg, Segalowitz & Carter, 2004; p. 135)

> Voluntary action implies a subjective experience of the decision and the intention to act […] For willed action to be a functional behavior, the brain must have a mechanism for matching the consequences of the motor act against the prior intention. (Sirigu et al. 2004; p. 80)

So it does seem as if the Comparator plays a substantial role in many theories. Just looking at the quotations above, to be able to compare the goals with the outcome of one's actions is deemed of vital importance for as diverse things as rational decision making, learning and voluntary action. The Comparator should have been on full alert when the choice was executed.

To connect with the discussion of Fodor's version of Intentional Realism, a third alternative is of course that there is nothing in the box. Or, at least, whatever process fills the role of initiating the action or representing the desired goal-state, it does not correspond to what could be expected from a Folk Psychological perspective. The model still works if the only thing that is supposed to be represented is the motor action: I point, reach, and pick up the

picture to the right. That is what I did, so I did the right thing. But the model is clearly meant to be more than this. If I reach for a beer but end up with a glass of milk in my hand, I should notice, because that was not what I wanted! In relation to standard models of goal-directed behaviour, I think choice blindness is a genuine problem that needs to be addressed.

A small caveat is called for here. I do not claim that it is impossible to consciously deliberate the reasons back and forth for a particular choice, and we certainly can remember (some) of the things we tell ourselves when doing so. And we can set up "goals" like quitting smoking and then notice when we fail to achieve them. But the things we *say* to ourselves when trying to quit smoking should not be the starting point when we try to build models for how our cognitive machinery represents the mechanisms for our actions. They are Folk Psychological constructions, given their exactness through the language we use, not by a reality they describe.

Introspection and verbal reports

Summary Paper 3: ***How something can be said about telling more than we can know.*** The experimental method in Paper 3 is identical to the one used in Paper 2. The participants were shown pairs of female faces and were asked to choose which one they found more attractive. After the choice had been performed, the participants were sometimes asked to explain their choice. Eighty participants completed 15 trials each, of which three were manipulated. The deliberation time for performing the choice was fixed to four seconds for all conditions. The set of faces was different from that used in Papers 1 and 2.

The important difference in relation to the study described in Paper 2 is the collection of introspective verbal reports. This study was divided into two different conditions. In the first condition the participants were simply asked why they preferred the chosen picture. The same question was asked in the second condition, but now the experimenter encouraged the participants to elaborate their answers up to one full minute of talking time. This was done both by the use of positive verbal and non-verbal signals and by interjecting simple follow-up questions.

Two major methods were used in the comparative analyses of the verbal reports: relative word frequency and latent semantic analyses. Based on relevant research, such as automatic lie-detection and

> language development, a large number of variables were compared for manipulated and non-manipulated reports. Examples are: filled and unfilled pauses, words marking uncertainty, specific and non-specific nouns, positive and negative adjectives, lexical density and diversity. Of the total 30 variables measured for long as well as short reports, only two variables were statistically different in manipulated and non-manipulated reports. In latent semantic analyses, by analysing the contextual usage of words in a large corpus (i.e. a collection of text), a "semantic space" is constructed representing the relative distance between the words in the corpus. This space can in turn be used to calculate the difference between two other corpora. In our analyses, we found no difference between manipulated and non-manipulated reports. In contrast, large discrepancies were found between our male and female participants, both with latent semantic analyses and with several of the linguistic frequency variables. The detection of sex differences shows that it is possible to detect differences in our corpus with the methods we have used, which thereby gives strength to the overall conclusion that there are very few differences between manipulated and non-manipulated reports.

No difference that makes a difference

To better appreciate the discussion of the theoretical context of this study, a few words on the underlying reason for examining the verbal reports. First of all, it is interesting that the participants *do* talk in the manipulated trials, that they say anything at all. As they are asked to explain a choice they did not make, saying "I don't know" or "I wanted the other one!" would seem the more natural thing to do.

Secondly, it is interesting to analyse what the participants actually say, to find out to what extent they give reasons referring to the original choice or the manipulated outcome. Due to the nature of the stimuli it is often quite hard to determine which of the faces has the "pretty nose" or the "nice haircut" the participants might claim to have been influential in their decision. But sometimes the features referred to are unique for the manipulated picture, such as the earrings, the dark hair or a hint of a smile. In these cases we can be certain that the reports are constructed after the fact, and thus in some sense are confabulatory.

Thirdly, it is interesting to compare the manipulated and the non-manipulated reports. The amount of difference detected says

something about the "normality" of the manipulated reports. If the reports have the same amount of detail, the same number of pauses and markers of uncertainty, the same amount of emotional content, and so on, then there is nothing "wrong" with the reports generated in the manipulated trials. This also serves as an implicit marker whether the participants on some unconscious level have detected or registered the manipulation, as detection might have asserted itself, for example, in an increase of markers of uncertainty.

And finally, the lack of differentiation between manipulated and non-manipulated reports also says something about the "authenticity" of the non-manipulated reports. If there are no or few differences between manipulated and non-manipulated reports, and we know that the manipulated reports at least to some extent are confabulatory, then this might indicate that the same mechanism is responsible for both types of reports. In this roundabout way, it could be argued that the problems of finding differences between manipulated and non-manipulated reports are due to the fact that they are *both* confabulatory. No difference that makes a difference.

Know Thyself

Hardly any concept in the history of psychology and philosophy of mind has generated more controversy than introspection (Lyons 1986). Since Descartes' dualist vision of a mind fully transparent to the self, the pendulum regarding just how much we think we know about ourselves *from the inside* has swung back and forth several times. Early experimental approaches such as the German Gestalt psychology (e.g. Wertheimer, 1912) relied heavily on the ability to report accurately on one's perceptual experiences. This was in turn followed by Methodological Behaviourism (Watson, 1913; Skinner, 1938), in which behaviour is supposed to be explicable without reference to intermediate mental states, leaving little interest for what people claimed to know about the workings of their own minds. Despite not being necessarily committed to introspection, the cognitive movement that came to replace Behaviourism as foundational for psychological research at least put the mental back on the map. Still, prominent researchers such

as Ericsson and Simon (1980; 1998) believe that by using techniques such as "think aloud" during problem solving, we get an accurate picture of what is actually going on when we make decisions and solve problems.

A parallel and not entirely coincidental development can be seen in philosophy. In the phenomenological tradition, Husserl developed the notion of *epoché*, which translates to an isolation of the inner experience from theories or preconceptions of how the world works. The subjective perspective is essential for understanding the mind, and the goal to strive for is the "purest" form of introspection (Husserl, 1900/1970). Wittgenstein questioned this very idea in his famous discussion of the private object (Wittgenstein, 1953). It is of course not entirely clear what Wittgenstein would recommend as psychological practice, but he is at least often interpreted as arguing against the possibility of isolating an experience and then saying something meaningful about it. In *Concept of Mind*, Ryle (1949) was a bit more straightforward in his attack on introspective knowledge:

> The sort of things that I can find out about myself are the same as the sort of things that I can find out about other people, and the methods of finding them out are pretty much the same. A residual difference in the supplies of the requisite data makes some difference in degree between what I can know about myself and what I can know about you, but these differences are not all in favor of self-knowledge. (p. 155)

For Ryle, mental talk was to be understood as dispositions to act, not as descriptions of causally active entities. Despite being out of favour nowadays, Ryle's Logical Behaviourism inspired many later thinkers, such as Sellars (1963) and Dennett (1987; 1991a).

In modern days, the debate over the use and utility of introspection has been seamlessly intertwined with the discussion of the "easy" and the "hard" problems of consciousness (Chalmers, 1996) that is, what can be known about consciousness from the first person perspective (introspection) compared to the third person perspective (the standard scientific method).

The partisanship is as fierce as ever concerning the philosophical problems of consciousness and introspection, with cemented positions and slight chances of resolution or reconciliation (Block, 1995; Chalmers, 1996; Dennett, 1993; Rorty, 1993).

In cognitive science, something like a consensus has emerged around a picture of the mind as primarily being made up out of unconscious machinery (e.g. see Gazzaniga, 2004). It is clear that large parts of what is going on in the brain do not ever reveal themselves to introspection (Dehaene & Naccache, 2001; Wilson, 2002; LeDoux, 1996). But there is also a steadily growing appreciation for the central role introspective reports can play in, for example, cognitive neuroscience research, triangulating the reports with behaviour and brain activity (Jack & Shallice, 2001; Jack & Roepstorff, 2002).

There are many forms and aspects of introspection, as there are many different things we can know about ourselves, our experiences and our mental states (Schwitzgebel, 2002). To lump all threads together in one quick historical sweep does not do justice to the intricacies of all positions held and argued for. For example, in relation to phenomenal states or qualia (things like seeing red or the softness of a kiss), I cannot claim that our experiments have much to say. Regarding self-knowledge and introspection as such, I am primarily concerned with higher-order mental states such as beliefs and desires. And in relation to this, what introspection can tell us about what we believe and what we desire, our experimental results clearly support an anti-introspectionist view. If we are supposed to know our own minds from the inside, we should know why we do what we do. And when asked to describe why we chose a face we in reality did not prefer, we are not supposed to just fabricate reasons (at least not without knowing that this is what we are doing). In our experiments, it is evident that the participants do not have perfect access to their underlying cognitive machinery. But despite being a striking demonstration that we don't always know why we do things, the results of our experiments do not have as great an impact on philosophy of mind as they might have had some decades back. Few philosophers today believe us to be infallible concerning our own mental processes. However, in relation to the previously mentioned debate about how we use Folk Psychology, introspective knowledge *is* essential for philosophers such as Goldman (1993), as we are supposed to understand the behaviour of other people through an internal simulation of what we would have believed and desired

had we been in their shoes. If we use our own mind as a model to understand others, it is a bit curious that we have such a lack of understanding of how we function ourselves.

Regardless of what we actually *do* know about our own mental life, one interesting aspect of self-knowledge is that for most people it does *feel* as if we know ourselves from the inside.

As in our case, when I tell you why I made a particular choice, I just assume that I am right. Where this sense of knowing comes from is of course contested (i.e. does it feel right because in general we *are* right? Goldman, 1993; Gopnik, 1993), but most people debating introspection agree that this is a prevalent part of the psychological sphere. One reason why it feels as if we have this special authority about ourselves is that we are very seldom proven wrong. However strongly I suspect that "being sorry" does not accurately describe your present condition, when you tell me that this is how you feel, there is no external evidence for me to use against your claim. But this is true in relation to ourselves as well, i.e. we rarely realise that we are wrong in our self-explanations. As Nisbett and Wilson (1977; p. 256) say: "disconfirmation of hypotheses about the workings of our own minds is hard to come by." This is also a genuine problem when doing experimental work on self-knowledge. Without any means to question the validity of people's verbal reports, it is also difficult to say how much of it is true. Most often, the correctness of people's introspective reports is just taken for granted.

We have solved this problem in our experiment. We do not need to take on the burden of explaining the mechanism behind the original choice – why they preferred one face over the other in the first place. Given the structure of the manipulation, we just *know* that the participants did not want what they got. By setting up this mismatch between what they wanted and what they received, we now have a way of demonstrating when experimental participants are manifestly wrong about themselves. And as such it is a novel tool in research on self-knowledge. And in addition, it is also a way to show both to ourselves and to others that we do not know as much about ourselves as we think we do.

As was the case in the previous discussion of choice blindness and Folk Psychology, the implications of our approach are per-

haps better seen when we connect it to a more specific research tradition in cognitive science.

How something can be said

Paper 3 takes as its starting point the classic article "Telling More Than We Can Know: Verbal Reports on Mental Processes" by Nisbett and Wilson (1977). It is one of the most cited articles of all times in psychology as well as philosophy, and it surfaces in the most diverse circumstances. But what did they actually say to stir such a controversy?

At the outset, Nisbett and Wilson make clear that they are interested in mundane verbal interactions, such as giving and taking reasons, asking questions, making judgements, stating preferences, etc. In our daily lives, we are confronted with countless questions that rely upon our higher-order cognitive processes: "Why do you like him?" "How did you solve this problem?" "Why did you take that job?" (1977, p. 232). We answer such questions with apparent ease, and we ask them ourselves believing that others can tell why they do what they do. Nisbett and Wilson thought this confidence ill-founded. They had collected a lot of relevant research from neighbouring fields, as well as performing a large number of experiments themselves. Their own (rather harsh) verdict:

> [T]here may be little or no direct introspective access to higher order cognitive processes. […] when people attempt to report on their cognitive processes, that is, on the processes mediating the effects of a stimulus on a response, they do not do so on the basis of any true introspection. Instead, their reports are based on a priori, implicit causal theories, or judgments about the extent to which a particular stimulus is a plausible cause of a given response. (Nisbett & Wilson, 1977; p. 232)

They had reviewed large parts of the then burgeoning experimental social psychology literature, with topics such as cognitive dissonance, insufficient justification, and attribution theory, and found a lot of support for their conclusion.

An example of the kind of studies they leaned on is Zimbardo's famous grasshopper experiment (Zimbardo et al, 1969). This study is also a nice illustration of the insufficient justification effect, as well as a telling example what was allowed before the reign of ethics committees.

The group of participants consisted of students recruited to an outdoor survival training course. Naturally, to survive outdoors, an essential skill is learning to eat what nature has to offer. On this topic, how to best capture, prepare and eat grasshoppers was explained to the participants. Half of them were instructed by a nice and warm person, sensitive to their discomforts, interacting in a friendly manner with his assistants, etc. The other half were given an angry and hostile instructor, yelling at his co-workers, laughing at the participants, and so on. After the "eating" was done, the participants had to indicate what they actually thought of the experience. In line with insufficient justification theory, the group with a non-pleasant instructor liked the taste better than the other group (a few even took extra grasshoppers home to share with their friends and families). The logic of insufficient justification theory is sometimes a bit hard to follow, but to explain using the terms of the theory: In the first group, the "dissonance" between disliking grasshoppers and still eating them could be reduced by "thinking" that they did it because the instructor was such a nice man, and as the dissonance was accounted for by referring to the instructor, the participants did not need to change their negative attitudes towards eating grasshoppers. But in the second case, the participants could not find a sufficient justification for why they ate those disgusting grasshoppers, so they changed their attitude towards liking them instead. It is the same argument as in experiments in which you like a boring task more if you get paid less; as it can not have been the money that made you do it, you must just have liked it!

But what is important here is that the participants themselves are not aware that their attitude has been influenced by the behaviour of the experimenter. If asked why they would not have known that the perceived likeability of the instructor was the reason they now (believed themselves) to like eating grasshoppers.

Among Nisbett and Wilson's own experiments, the most pertinent to our experiments is the stocking and nightgown study. Under the pretence of a consumer survey, people walking by in a shopping centre were invited to evaluate articles of clothing. The participants were either asked to indicate which one of four different nightgowns they preferred, or to evaluate four identical

pairs of nylon stockings. When they had made their choice, they were asked why they had chosen the article in question. As reported by Nisbett and Wilson: "there was a pronounced left-to-right position effect, such that the right-most object in the array was heavily overchosen. For the stockings, the effect was quite large, with the right-most stocking being preferred over the leftmost by a factor of almost four to one" (1977, p. 243). In contrast to this, none of the participants mentioned position as having a possible influence on their choice; not surprisingly, they commented on the quality or texture of the fabric instead. Nisbett and Wilson themselves were not able to provide a systematic explanation of why position should be such an important factor. Their suggestion was that people might examine the items from left to right and hold of judgement until the last one in the array had been explored. But what is important here is not really *how* the ordering influenced the evaluation, the interesting part is that we know that it had an effect but still did not show up in the participants' own explanations.

The stocking and nightgown study nicely captures the spirit of the Nisbett and Wilson approach, showing that we sometimes are unaware of which stimulus influences our behaviour. It is also relevant because it bears a structural resemblance to our studies: several items are evaluated, one of them is publicly chosen as the one preferred, and the choice is later explained to the experimenter. But there are also some important differences. Naturally, I consider our choice blindness experiments to represent a methodological step forward. By listing some of the arguments directed against the studies of Nisbett and Wilson, we can see to what extent that is true.

Ecological validity. Nisbett and Wilson have been accused of using unimportant and contrived tasks in their experiments: It is somewhat strange to choose the one preferred of *identical* stockings at a clothing retailer (Kraut & Lewis, 1982; Kellogg, 1982). It is not unreasonable to believe that our introspective capacities may be diminished under such circumstances (Smith & Miller 1978). In contrast, choosing which face one finds more attractive is a very straightforward task, reflecting a simple type of judgement that people often make in their daily lives. While not being the

most important task imaginable, many people have very strong opinions about facial attractiveness. Compared to the studies of Nisbett and Wilson (and to psychological experiments in general), evaluating faces is as interesting as it gets.

Verbal reports. Despite the title of their article, very little was done with the verbal reports in Nisbett and Wilson (1977). Apart from registering whether the influential stimuli were mentioned or not, no thorough or comparative analyses were performed. In most of the experiments the introspective reports were also generated several minutes (or even hours) after the critical behaviour occurred. Several critics therefore argued that the impoverished and "incorrect" verbal reports were due to a memory effect (Ericsson & Simon, 1980). The participants had simply forgotten why they did what they did. Ericsson and Simon (1980; 1993) put this in contrast to their own protocol analyses and "think-aloud" technique, in which the participants "reveal" their actual trains of thought by verbally stating what they think while performing a task. If done properly, with the correct timing, this is supposed to yield a "correct" description of our cognitive processes:

> [T]he validity of verbally reported thought sequences depends on the time interval between the occurrence of a thought and its verbal report, where the highest validity is observed for concurrent, think aloud verbalizations. For tasks with relatively short response latencies (less than 5–10 seconds), subjects are able to recall their sequences of thoughts accurately immediately after the completion of the task and the validity of this type of retrospective reports remains very high. (Ericsson, 2002; p. 3)

In our experiments, the reports were solicited only a few seconds after the choice was made, immediately after the participants had received the chosen picture. According to the quotation above, this is well within the time margin Ericsson has set up for delivering accurate descriptions of our cognitive processes. What the participants say in our experiment should be a true reflection of why they chose one picture over the other. In a way Nisbett and Wilson's studies did not, our results seem to challenge this position.

It should also be noted that in our experiments the participants had been informed at the beginning of the sessions that we would ask them about their reasoning, thus cueing them to reason deliberately, and to attend to their reflective processes.

Individual vs. group effects. In most of the experiments presented in Nisbett and Wilson (1977), the discrepancies between action and introspection can only be discerned in group-level response patterns, not for each individual (Quatrone, 1985; Quatrone & Jones, 1980, Smith & Miller, 1978; White, 1988). In the stocking and nightgown experiment above, it is impossible to say which of the participants were influenced by the positioning of the items, we only know that some of them must have been influenced as we know that from a statistical perspective there is an ordering effect. In our experiments, we *know* that the participants did not want the photograph received in the manipulated trials. Whatever the participants say, it will be in contrast to what they originally intended to choose. This design also gives us the two classes of verbal reports to compare and contrast. And at the very minimum, in the manipulated reports describing unique features of the non-chosen picture, we have unequivocally shown that normal participants may produce confabulatory reports when asked to describe the reasons behind their choices. This too goes beyond what was established by Nisbett and Wilson.

I think we are allowed to say that our experiment is a methodological improvement on what was employed by Nisbett and Wilson. We solve several of the problems they were criticised for, as well as providing a methodological platform for new experiments. Our experimental design is the first to give cognitive scientists the opportunity to systematically study how confabulatory reports are created and how they relate to standard or "truthful" reports about choice behaviour. In the end, this will hopefully enable us to also say something about the *general properties* of introspective reports.

Methods and Methodology

Summary Paper 4: ***Magic at the marketplace.*** The experiment took place inside a local supermarket, and the participants were recruited after being asked if they wanted to participate in a consumer preference test. The test consisted of tasting or smelling two sorts of jam and two sorts of tea. When the participants had made their choice of which jam or tea they preferred they got to sample the chosen item again, and were asked to explain why they liked this one better. For each participant, either the tea or the jam condition was manipulated. By using prepared jars with two separate compartments containing both varieties of jam or tea, the experimenter could switch the position of the two jams or teas by simply turning both jars upside down (see Figure 1 in Paper 4). When the participants sampled the third time they were given of the non-chosen product, and at the same time they were asked why they liked this taste or smell better,

In total, 180 participants took part in this experiment. The similarity within the pairs was established in a pilot study. The six pairs used in the experiment ranged from relatively similar to distinctively dissimilar. A trial was categorised as detected if the participants voiced any concerns immediately after tasting or smelling the switched jam or tea or if the participants at the end of the experiment in any way claimed to have noticed the manipulation. A manipulated trial was also considered detected if the participant thought that the taste or smell had changed the second time it was sampled.

Half the participants also received either a package of tea or a jar of jam as a gift. The jam or tea chosen by the participants in the manipulated trials was also the product used as gift. In addition, several other factors were measured in the experiment. When sampling the first time, the participants rated both sorts of jam and both sorts of tea with regard to how good they tasted or smelled. After the choice, the participants rated how easy it was to discriminate between the two choice options, and also indicated how confident they were in their choice.

Counting all conditions and all forms of detection, 32.2% of the manipulated tea trials and 33.3% of the manipulated jam trials were detected. There was an increased rate of detection for the least similar compared to the most similar pair for both tea and jam. The gift was associated with lower detection rate for tea but not for jam. A larger discrepancy in attractiveness rating was associated with higher degree of detection for jam but not for tea. Comparing manipulated and non-manipulated trials, the perceived ease of distinguishing between the items in the pairs was higher for non-manipulated trials for tea but no difference was found for jam. There were no differences in rated confidence between the manipulated and the non-manipulated trials for either tea or jam.

> The major conclusion drawn is that choice blindness is further established as a robust effect in decision making, extending the findings from previous research using visual stimuli to the modalities of taste and olfaction.

The Wedge

At the beginning of the introduction, I identified three things as novel in this thesis: Choice blindness, the verbal reports and the experimental methodology as such. The first two entries on this list have been discussed in relation to Papers 1, 2 and 3. Accordingly, Paper 4 will be primarily used as a platform for a discussion of the experimental methodology. I will give some background for why and how we came up with the idea of doing the kind of studies described in the thesis, and also present some planned future work on choice blindness.

From a methodological perspective, it is important to point out that the experimental approach was deduced from our theoretical background rather than the other way around, i.e. we did not invent the experiments first and then try to find a suitable context for them. Being very much influenced by Daniel Dennett, my colleague Lars Hall and I had for a long time thought that there must be some experimentally testable consequences of his Intentional Stance theory. We had previously made a distinction between (the classical concept of) introspection and a more Dennettian mode of self-knowledge based on self-observation, which we called *extrospection* (Hall 2003, Hall & Johansson 2003a). To emphasize the potential of extrospection as a tool for self-understanding, we had applied this concept in the domains of educational psychology and self-control (Hall & Johansson 2003b, Hall, de León & Johansson, 2002), but thus far we had not made a direct empirical test of the theory.

Given this perspective it ought to be possible to influence people's interpretations of themselves by controlling what evidence they have available for their extrospective reasoning. As Dennett claims in the long quotation I used previously, every one of us is an: "inveterate auto-psychologist, *inventing* intentional interpretations of our own actions in an inseparable mix of confabulation, retrospective self-justification, and (on occasion, no doubt)

good theorizing" (Dennett, 1987; p. 91, emphasis in original). As we see it, choice blindness can be used as a wedge to pry apart the otherwise "inseparable mix" of the things we do and the things we say about ourselves.

An interesting further application of this methodology is to examine what happens *after* the choice (what Dennett 1991a calls The Hard Question: *And then what happens?*). In Paper 1, a memory test used after the completion of the choice experiment revealed that the participants tended to remember the manipulated outcome as being what they originally preferred. But the more interesting question is what becomes of the participants preferences and attitudes; what would for instance happen if they had to do the same choice again, would they pick the alternative they initially thought was better or the mismatched option they ended up with?

We have recently begun to explore this question. In the experiment that formed the basis for the introspective reports that were analysed in Paper 3, the participants had to choose between two faces, pick the one they preferred, and give either a short or a long verbal report explaining their choice. But in addition to this, their later preferences were also probed in several different ways. All participants were presented with the pairs a second time and had to choose the picture preferred once again. In one condition, the participants also had to rate on a numerical scale how attractive they thought both pictures were directly after having given their verbal reports. The results showed that the participants were clearly influenced by the manipulations made, as they were much more likely to pick the originally non-preferred face the second time they had to evaluate a pair. But perhaps even more interestingly, this tendency was correlated with the participants "involvement" in the choice, i.e. if they had given short or long reports, and if they had numerically rated the pictures after the first choice (see Hall, Johansson, Tärning & Sikström, in preparation). [4]

We think this is a very interesting avenue of exploration. What will happen with these "induced" preferences over time? Will they

4. This paper was meant to be included in the thesis, but life could no longer wait.

transfer to more general attributes (like preferring brunettes)? Will they be modulated by other choices? In a sense, choice blindness can be used as an instrument to measure how much we influence *ourselves* by the choices we make.

A brief note on magic

The experimental procedure in Paper 2 was developed in cooperation with the eminent Swedish close-up magician Peter Rosengren. The technique used is called "black art" (Dondrake, 2003), which is a method of concealing something black against a black background (e.g. the ropes carrying the attractive assistant when she appears to float in mid-air on stage). In the manipulated trials, the experimenter held two cards in each hand, with the card shown fitted with a black back side of the same material as the black desk cover that served as the surface of the experiment. When the "chosen" picture was slid to the participant, the front card stayed on the table. Generally, black art can be used effectively even at a very close range, but since we needed to conduct our experiment in a brightly lit office environment we also used some sleight of hand, through which the extra card is hidden by the experimenter's sleeve until it is raked back and falls down in a hidden compartment at the end of the table (see Picture 1 in Paper 2).[5]

The technique used in Paper 4 has its origin in a long discussion we had with two professional magicians at the yearly "Swedish Magicians' Circle" conference, Karl Berseus and Axel Adlercreutz. But it was Lars Hall who came up with the brilliant idea of gluing two jars together and thereby creating a single jar with two separate compartments. In this experiment, we also used two experimenters working together to conceal the manipulation, as the first experimenter waits to execute the switch until the participant moves his or her attention to the other experimenter to answer a question about how well they liked the sampled item.

Interestingly, while the techniques of the experiment are imported from the domain of magic, the purpose of the experiments is more or less the opposite of what magicians usually want to

5. In the experiment in Paper 2, only two participants were removed for having seen the procedure – as they would say in the classic poker movie *Rounders*: I only got caught with a hanger twice!

achieve. In card magic, the performer must take great pains to ensure that the participants and the members of the audience are able to remember which card was initially chosen. Otherwise, when the act reaches its finale, they would simply be unable to notice that anything magical had taken place. But in our experiments the whole point is the participants not noticing the change; in this case, we have to wait for the applause until we are published! Despite this, it is safe to say that it has been a lot more fun to invent and perform the experiments than to analyse the data obtained.

The future of choice blindness

As we see it, there are a great number of possible variations and extensions that can be made in relation to the experiments we have produced so far. In both Paper 1 and Paper 4, we briefly discuss the possibilities of using the methodology of choice blindness as a more general tool in psychological research. Here, I would just like to give a short overview of some of the things we have started on or plan to do in the near future.

We do not yet know the limits of choice blindness. For instance, while it seems as if it would be impossible to swap two pictures of Marilyn Monroe and Marilyn Manson without the participants noticing, it is still an empirical question how dissimilar or how "unequal" two pictures can be. We also need to investigate more rigorously the importance of parameters related to the memory of the choice, such as the encoding time (i.e. the time participants are allowed to deliberate upon their choice), the occlusion interval (i.e. the time the chosen stimuli is invisible when the manipulation is performed), and the retention interval (i.e. the time until the mismatch detection is tested).

But we can also change the stimuli as well as the task to be performed by the participants. Both abstract patterns and male and female faces have been tested, but perhaps change blindness would disappear if other stimuli were used (as someone remarked in an Internet chat-forum after the *Science* publication: "Who cares about pictures of young women – had it been pictures of new cars there is no way I would have missed the switch!"). We could also use more "culturally" charged stimuli, such as brands

or logotypes (fake or real), and ask question in line with standard marketing research, for instance, which symbol is more energetic, youthful, dynamic etc. If we keep faces as stimuli but instead change the task, we could, for example, vary the importance of the choice, such as letting the participants choose which of two persons they were going to have a cup of coffee with, or which one they would prefer to employ at their company.

Large parts of the research done on face processing have been on aspects relevant from an evolutionary perspective, and much of this research is easily adapted to our approach (Penton-Voak, & Perrett, 2000; Perrett et al., 1999). For example, we could systematically vary the symmetry of the faces, or change the task to things like which person would you rather have a long-term relationship with as compared to a one-night stand. It could be suspected that changes made on more evolutionarily important choices should also be more easily detected, but again, this is an empirical question.

To expand on the issue of verbal reports and confabulation, instead of a complete identity switch, we could just add potentially salient features, such as earrings or a smile, and see if any of these features were mentioned in the participants' explanations of their choice. If they were, this would add even more strength to the suspicion that reasons stated for choices are often constructed "after the fact". But there are no grounds for *not* including verbal reports in all or most of the experiments, and thereby building a large "database" of various forms of manipulated and non-manipulated reports.

One large class of data that we have yet to work with is implicit measures, such as galvanic skin response, eye-tracking, ERP and fMRI. This type of measures is interesting for several different reasons. First of all, they might reveal specific response patterns that differentiate between manipulated and non-manipulated trials, indicating that, despite the participants' own conscious denial of having detect a manipulation, some parts of the cognitive system actually "noticed" that something went wrong with the choice. There is a large literature on change blindness and change detection in general that is connected to this issue (see Simons & Silverman, 2004). Secondly, there might be patterns in, for ex-

ample, the saccadic movement of the eyes that are indicative of whether a change *is going to be* detected. Perhaps the detected manipulations are encoded differently? Thirdly, there might be ways to connect the verbal reports to, for example, patterns revealed by ERP. Are there any differences in activity between giving confabulatory and "ordinary" verbal reports?

By keeping the methodology and just varying the stimuli and the task, a large number of interesting experiments could be made. But we could also expand on the method, using new "magic" tricks, such as the prepared jars in Paper 4. With methods like this we could try changing real objects rather than just pictures, as well as further exploring choices in other sensory modalities than vision.

As suggested by the inclusion of implicit measures, we can also focus on other aspects of the participants' responses. One interesting (and underdeveloped) feature in Paper 4 is the certainty measure – i.e. the participants' own rating of how certain or confident they felt in their choice. We found no differences between manipulated and non-manipulated trials, which means that the participants were just as confident in a choice they did not intend to make as in one they did make without alterations. The use of self-rating scales of certainty is a prevalent component in psychological research on decision making (Baranski & Petrusic, 1998; Petrusic & Baranski, 2003; Pallier et al., 2002). The fact that it is possible to switch the outcome of people's choices without this making a mark on how confident they are in those choices ought to say something about the precision of this type of self-rating measures.

Similarly, the study in Paper 4 can be used as a starting point in a more thorough investigation of decision making and consumer behaviour. In what circumstances are we blind to changes in our consumer choices? How does a non-detected manipulation affect, for instance, how much we are willing to pay for a certain item, or how satisfied we are with a certain product after we have bought and used it? There are of course many other ways of working with choice blindness to illuminate previous research on choices and decision-making, as well as the use of "introspective" verbal reports in psychological research.

Another approach would be to further enquire into the participants' self-understanding in our experiments. What do they themselves think they do when they answer the question *why* they performed a choice; do they think they have access to their own psychological processes, or do they think they just report the most likely causes when looking at the picture the second time? How certain are they that what they say actually captures the reasoning process responsible for their decision? What would happen if we instead asked *how* they came to that conclusion – would they attempt a more causal account compared to a *why*-question, or would they say that they just don't know? The terms introspection and confabulation have a very special meaning in philosophical jargon, but what does it correspond to when laypersons try to describe themselves and the actions they perform?

Despite being a both brief and shallow run-through of some of the things on our to-do list in the next few years, I hope it has served the purpose of showing that choice blindness as a concept extends further than the four studies presented in the thesis.

The End of the Beginning

There are of course many more things I would like to say in relation to my thesis. But it is time to stop here and let the papers talk for themselves.

As a final note, I would like to point out that even if this is my thesis, the work behind it is very much a collaborative effort. Lars Hall and I have worked on this project for a very long time, and during the last few years our duo has turned into a full group. Therefore, I would like to share the credit with all those listed as co-authors on the papers, but take the blame myself for all faults to be found herein.

References

Astington, J. W. (1993). *The child's discovery of the mind*. Cambridge, MA: Harvard University Press.

Badre, D., & Wagner, A. (2004). Selection, integration, and conflict monitoring: Assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron, 41*(3), 473–487.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: experiments on the time to determine confidence. *Journal of Experimental Psychology Human Perception and Performance*, *24*(3), 929–45.

Baron-Cohen, S. (1994). How to build a baby that can read minds: Cognitive mechanisms in mindreading. *Current Psychology of Cognition, 13*(5), 513–552.

Baron-Cohen, S. (1995). *Mindblindness: an Essay on Autism and Theory of Mind*. Cambridge, MA: MIT-Press.

Block, N. (1995). On a Confusion of a Function of Consciousness. *Behavioral and Brain Sciences, 18*(2), 227–287.

Bogdan, R. J. (1991). (Ed.). *Mind and Commonsense Psychology*, Cambridge: Cambridge University Press.

Brentano, F. (1874/1973). *Psychology from an Empirical Standpoint*. New York: Humanities Press.

Brown, R. G., & Pluck, G. (2000). Negative symptoms: the 'pathology' of motivation and goal-directed behaviour. *Trends in Neurosciences, 23*(9), 412–417.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chisholm, R. (1966). *The Theory of Knowledge*. Englewood Cliffs, NJ: Prentice Hall.

Christensen, S., & Turner, D. (1993). (Eds.). *Folk Psychology and the Philosophy of Mind*. New York: Lawrence Erlbaum Associates.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, *78*, 67–90.

Dehaene, S., & Naccache, L. (2001). Toward a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*, 1–37.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dennett, D. C. (1991a). *Consciousness explained*. Boston: Little, Brown & Company.
Dennett, D. C. (1991b). Real patterns. *Journal of Philosophy*, *89*, 27–51.
Dennett, D. C. (1993). Back from the Drawing Board. In B. Dahlbom (Ed.). *Dennett and his Critics: Demystifying Mind*. Oxford: Blackwell.
Drake, D. (2003). Dondrake's Black Art Breakthroughs. Dondrake.
Ericsson, K. A., & Simon, H. A. (1980). Verbal Reports as Data. *Psychological Review, 87*(3), 215–251.
Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, *5*(3), 178–186.
Ericsson, K. A. (2002). Protocol analysis and verbal reports on thinking. Retrived 15 September 2006, from: .http://www.psy.fsu.edu/faculty/ericsson/ericsson.proto.thnk
Fodor, J. A. (1983). *The modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.
Fodor, J. A. (2000). *The mind doesn't work that way: The scope and limits of computational psychology*. Cambridge, MA: MIT Press.
Gazzaniga, M. S. (2004). (Ed.). *The New Cognitive Neurosciences III: Third Edition*. Bradford Books.
Goldman, A. (1993). The psychology of folk psychology. *Behavioral and Brain Sciences*, *16*, 15–28.
Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, *16*, 1–14.
Greenwood, J. D. (1991) (Ed.). *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press.
Hall, L. (2003). *Self-Knowledge/Self-Regulation/Self-Control: A Ubiquitous Computing Perspective*. Ph.D. Thesis. Lund University Cognitive Studies, 113.

Hall, L., & Johansson, P. (2003a). Introspection and Extrospection: Some Notes on the Contextual Nature of Self-Knowledge. *Lund University Cognitive Studies, 107*.

Hall, L., & Johansson, P. (2003b). Self-Regulation in Education: A Ubiquitous Computing Perspective. *Lund University Cognitive Studies*, *111*.

Hall, L., de Léon, D., & Johansson, P. (2002). The Future of Self-Control: Distributed Motivation and Computer-Mediated Extrospection. *Lund University Cognitive Studies, 95*.

Hall, L., Johansson, P., Tärning, B., & Lind, A. (in press). How Something Can Be Said About Telling More Than We Can Know: Reply to Moore and Haggard. *Consciousness and Cognition*.

Hall, L., Johansson, P., Tärning, B., & Sikström, S. (in preparation). Choice Blindness and Preference Change. *Lund University Cognitive Science*.

Husserl, E. G. A. (1900/1970). *Logical Investigations*. Vol 1-2.

Ingvar, D. (1999). On volition: a neurophysiologically oriented essay. In B. Libet, A. Freeman, and K. Sutherland (Eds.). *The Volitional Brain*. Exeter: Imprint Academic.

Jack, A. I., & Roepstorff, A. (2002). Introspection and cognitive brain mapping: From stimulus-response to script-report. *Trends in Cognitive Sciences*, *6*, 333–339.

Jack, A. I., & Shallice, T. (2001). Introspective physicalism as an approach to the science of consciousness. *Cognition*, *79*, 161–196.

Jeannerod, M. (2003). Consciousness of action and self-consciousness. A cognitive neuroscience approach. In J. Roessler, and N. Eilan (Eds.). *Agency and self awareness: Issues in philosophy and psychology*. Oxford: Oxford University Press.

Kellogg, R. T. (1982). When Can We Introspect Accurately About Mental Processes. *Memory & Cognition, 10*(2), 141–144.

Kraut, R. E. and Lewis, S. H. (1982). Person Perception and Self-Awareness - Knowledge of Influences on Ones Own Judgments. *Journal of Personality and Social Psychology, 42*(3), 448–460.

LeDoux, J. E. (1996). *The emotional brain*. New York: Simon and Schuster.

Lyons, W. (1986). *The disappearance of introspection*. Cambridge, MA: MIT Press.

Moore, J. W., & Haggard, P. (in press). Commentary on "How Something Can Be Said About Telling More Than We Can Know: On Choice Blindness and Introspection". *Consciousness and Cognition.*

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Pallier, G, et al. (2002). The role of individual differences in the accuracy of confidence judgments. *Journal of General Psychology*, *129*(3), 257–99.

Penton-Voak, I. S. and Perrett, D. I. (2000.) Female preference for faces change cyclically: Further evidence. *Evolution and Human Behaviour, 21*, 39-48.

Perrett, D. I., Burt, D. M., Penton-Voak, I. S., Lee, K. J., Rowland, D. A., and Edwards, R. (1999.) Symmetry and Human Facial Attractiveness. *Evolution and Human Behaviour, 20*, 295-307.

Petrusic, W. M., & Baranski, J. V. (2003). Judging confidence influences decision processing in comparative judgments. *Psychon Bull Rev, 10*(1), 177–83.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Quattrone, G. A. & Jones, E. E. (1980). Perception of Variability within in-Groups and out-Groups - Implications for the Law of Small Numbers. *Journal of Personality and Social Psychology, 38*(1), 141–152.

Quattrone, G. A. (1985). On the Congruity between Internal States and Action. *Psychological Bulletin, 98*(1), 3–40.

Quine, W. V. (1960). *Word and Object*. Cambridge, MA: MIT Press.

Ridderinkhof, K. R., Ullsberger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The Role of the Medial Frontal Cortex in Cognitive Control. *Science*, *306*, 443–447.

Ridderinkhof, K. R., van den Wildenberg, W. P. M., Segalowitz, S. J., & Carter, C. S. (2004). Neurocognitive mechanisms of cognitive control: The role of prefrontal cortex in action selection, response inhibition, performance monitoring, and reward-based learning. *Brain and Cognition*, *56,* 129–140.

Rorty, R. (1993). Holism, Intrinsicality, and the Ambition of Transcendence. In B. Dahlbom (Ed.), *Dennett and his Critics: Demystifying Mind*. Oxford: Blackwell.

Rushworth, M. F., Walton, M. E., Kennerley S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Science*, *8*(9), 410–417.

Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.

Scholl, B. J., Simons, D. J., & Levin, D. T. (2004). 'Change blindness blindness': An implicit measure of a metacognitive error. In D. T. Levin (Ed.), *Visual metacognition in adults and children: Thinking about seeing*. Westport, CT: Greenwood/ Praeger.

Schultz, W. (2000). Multiple reward systems in the brain. *Nature Reviews: Neuroscience, 1*, 199–207.

Schwitzgebel, E. (2002). How well do we know our own conscious experience? The case of visual imagery. *Journal of Consciousness Studies, 9*(5–6), 35–53.

Searle, J. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Sellars, W. (1963). *Science, Perception and Reality*. London: Routledge and Kegan Paul.

Simons, D. J., & Silverman, M. (2004). Neural and behavioral measures of change detection. In L. M. Chalupa and J. S. Werner (Eds.). *The Visual Neurosciences*. Cambridge, MA: MIT Press. 1524–1537.

Sirigu, A., Daprati, E., Ciancia, S., Giraux, P., Nighoghossian, N., Posada, A., & Haggard, P. (2004). Altered awareness of voluntary action after damage to the parietal cortex. *Nature Neuroscience*, *7*, 80–84.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. New York: Appleton-Century.

Smith, E. R., & Miller, F. D. (1978). Limits on Perception of Cognitive-Processes - Reply to Nisbett and Wilson. *Psychological Review, 85*(4), 355–362.

Spence, S. A., & Frith, C. D. (1999). Towards a functional anatomy of volition. *Journal of Conciousness Studies. 6*, 11–29.

Walton, M. E., Devlin, J. T., & Rushworth. M. F. S. (2004). Interactions between decision making and performance monitoring within prefrontal cortex. *Nature Neuroscience*, *7*(11), 1259–1265.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, *20*, 158–177.

Wertheimer, M. (1912). Experimental studies of the perception of movement. *Zeitschrift für Psychologie*, *61,* 161–265.

White, P. A. (1988). Knowing More About What We Can Tell - Introspective Access and Causal Report Accuracy 10 Years Later. *British Journal of Psychology, 79*, 13–45.

Wilson, T. D. (2002). *Strangers to ourselves: Discovering the adaptive unconscious*. Cambridge, MA: Harvard University Press.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell Publishing.

Zimbardo, P., Weisenberg, M., Firestone, I., & Levy, B. (1969). Changing Appetites For Eating Fried Grasshoppers with Cognitive Dissonance. In P. Zimbardo (Ed). *The Cognitive Control of Motivation: The Consequences of Choice and Dissonance*. Glenview, IL: Scott Foresman, 44–54.