This is the final published version of:

Abstract: A fundamental assumption of theories of decision-making is that we detect mismatches between intention and outcome, adjust our behavior in the face of error, and adapt to changing circumstances. Is this always the case? We investigated the relation between intention, choice, and introspection. Participants made choices between presented face pairs on the basis of attractiveness, while we covertly manipulated the relationship between choice and outcome that they experienced. Participants failed to notice conspicuous mismatches between their intended choice and the outcome they were presented with, while nevertheless offering introspectively derived reasons for why they chose the way they did. We call this effect choice blindness.

For an overview of our research, and access to our publications, please see our Choice Blindness Lab page:

www.lucs.lu.se/choice-blindness-group/

to monitor baseline synaptic transmission (Fig. 4A). In wild-type slices, tetanic stimulation of pathway S1 (100 Hz, 1 s) evoked homosynaptic LTP together with a heterosynaptic depression of the neighboring S2 pathway. The addition of an A1 antagonist (DPCPX, 800 nM) prevented heterosynaptic depression (Fig. 4B). To control for effects of enhanced baseline transmission that result in the presence of DPCPX, we switched from normal artificial cerebrospinal fluid (ACSF) to one containing 2.4 mM Ca^{2+} and 0.6 mM Mg^{2+}, which enhanced synaptic transmission to $172 \pm 8.9\%$ ($n = 3$ slices) of that in control mice. We still found heterosynaptic depression to be intact ($64.9 \pm 8.2\%$, $n = 3$ slices) compared to ACSF controls ($72.0 \pm 8.6\%$, $n = 3$ slices). To determine whether astrocytes mediate adenosine-dependent depression, we repeated this study using dn-SNARE slices and found a virtual absence of heterosynaptic depression (Fig. 4C).

These studies place the astrocyte at center stage in the control of adenosine. Glial-released ATP, which is rapidly hydrolyzed to adenosine, leads to a persistent synaptic suppression mediated by A1 receptors. Because adenosine is implicated in the control of wake-to-sleep transitions (26, 27) as well as responses to hypoxia, the identification of the central role of the astrocyte in regulating this nucleoside offers mechanistic insights into these processes.

The kinetics of ATP hydrolysis and adenosine accumulation provide a synaptic network with unique spatiotemporal conditions to control synaptic transmission. Fast-acting synaptic transmitters such as γ-aminobutyric acid and glutamate have high-affinity uptake systems in the vicinity of the synapse that constrain the time and distance over which a transmitter acts. Synaptic activation of an astrocyte to release ATP removes these constraints, because it takes ~200 ms before adenosine begins to accumulate (28). This provides time for ATP diffusion to distant sites, where it depresses synaptic transmission through accumulated adenosine, thereby providing a mechanism for cross-talk to distant synapses. In addition to activity-dependent actions, astrocytes, by persistently suppressing excitatory synaptic transmission, enhance the capability of synapses to express synaptic plasticity. Thus, the integration of synaptic activity by the astrocyte leads to a widespread coordination of synaptic networks. By suppressing excitatory transmission, astrocytes regulate the degree to which a synapse may be plastic, and during the induction of LTP, astrocyte-derived adenosine depresses neighboring unstimulated pathways.

References and Notes

1. A. H. Cornell-Bell, S. M. Finkbeiner, M. S. Cooper, S. J. Smith, *Science* **247**, 470 (1990).
2. A. Verkhratsky, H. Kettenmann, *Trends Neurosci.* **19**, 346 (1996).
3. P. B. Guthrie *et al.*, *J. Neurosci.* **19**, 520 (1999).
4. V. Parpura *et al.*, *Nature* **369**, 744 (1994).
5. M. J. Schell, M. E. Molliver, S. H. Snyder, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 3948 (1995).
6. T. Fellin *et al.*, *Neuron* **43**, 729 (2004).
7. A. Araque, V. Parpura, R. P. Sanzgiri, P. G. Haydon, *Eur. J. Neurosci.* **10**, 2129 (1998).
8. J. Kang, L. Jiang, S. A. Goldman, M. Nedergaard, *Nat. Neurosci.* **1**, 683 (1998).
9. T. A. Fiacco, K. D. McCarthy, *J. Neurosci.* **24**, 722 (2004).
10. A. Araque, R. P. Sanzgiri, V. Parpura, P. G. Haydon, *J. Neurosci.* **18**, 6822 (1998).
11. A. Araque, N. Li, R. T. Doyle, P. G. Haydon, *J. Neurosci.* **20**, 666 (2000).
12. P. Bezzi *et al.*, *Nat. Neurosci.* **7**, 613 (2004).
13. L. Pasti, M. Zonta, T. Pozzan, S. Vicini, G. Carmignoto, *J. Neurosci.* **21**, 477 (2001).
14. Materials and methods are available as supporting material on *Science* Online.
15. Q. Zhang *et al.*, *J. Biol. Chem.* **279**, 12724 (2004).
16. H. Zimmermann, N. Braun, *J. Auton. Pharmacol.* **16**, 397 (1996).
17. T. V. Dunwiddie, B. J. Hoffer, *Br. J. Pharmacol.* **69**, 59 (1980).
18. T. V. Dunwiddie, S. A. Masino, *Annu. Rev. Neurosci.* **24**, 31 (2001).
19. M. Kukley, M. Schwan, B. B. Fredholm, D. Dietrich, *J. Neurosci.* **25**, 2832 (2005).
20. K. A. Moore, R. A. Nicoll, D. Schmitz, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14397 (2003).
21. M. A. Ackley *et al.*, *J. Physiol.* **548**, 507 (2003).
22. S. Coco *et al.*, *J. Biol. Chem.* **278**, 1354 (2003).
23. O. J. Manzoni, T. Manabe, R. A. Nicoll, *Science* **265**, 2098 (1994).
24. J. T. Porter, K. D. McCarthy, *J. Neurosci.* **16**, 5073 (1996).
25. J. M. Zhang *et al.*, *Neuron* **40**, 971 (2003).
26. T. Porkka-Heiskanen, L. Alanko, A. Kalinchuk, D. Stenberg, *Sleep Med. Rev.* **6**, 321 (2002).
27. R. Basheer, R. E. Strecker, M. M. Thakkar, R. W. McCarley, *Prog. Neurobiol.* **73**, 379 (2004).
28. T. V. Dunwiddie, L. Diao, W. R. Proctor, *J. Neurosci.* **17**, 7673 (1997).
29. We thank I. Levitan and T. Abel for constructive criticism on versions of this manuscript, T. Abel for his constant guidance, and R. Thresher, H. Bujard, and M. Brenner for constructs used to generate transgenic vectors. Supported by funds from the National Institute of Neurological Disorders and Stroke and the National Institute of Mental Health to K.M., S.J.M., and P.G.H.

Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task

Petter Johansson,[1]* Lars Hall,[1]*† Sverker Sikström,[1] Andreas Olsson[2]

A fundamental assumption of theories of decision-making is that we detect mismatches between intention and outcome, adjust our behavior in the face of error, and adapt to changing circumstances. Is this always the case? We investigated the relation between intention, choice, and introspection. Participants made choices between presented face pairs on the basis of attractiveness, while we covertly manipulated the relationship between choice and outcome that they experienced. Participants failed to notice conspicuous mismatches between their intended choice and the outcome they were presented with, while nevertheless offering introspectively derived reasons for why they chose the way they did. We call this effect choice blindness.

A fundamental assumption of theories of decision making is that intentions and outcomes form a tight loop (1). The ability to monitor and to compare the outcome of our choices with prior intentions and goals is seen to be critical for adaptive behavior (2–4). This type of cognitive control has been studied extensively, and it has been proposed that intentions work by way of forward models (5) that enable us to simulate the feedback from our choices and actions even before we execute them (6, 7).

However, in studies of cognitive control, the intentions are often tightly specified by the task at hand (8–10). Although important in itself, this type of research may not tell us much about natural environments where intentions are plentiful and obscure and where the actual need for monitoring is unknown. Despite all its shortcomings, the world is in many ways a forgiving place in which to implement our decisions. Mismatches between intention and outcome are surely possible, but when we reach for a bottle of beer, we very seldomly end up with a glass of milk in our hands. But what if the world were less forgiving? What if it instead conspired to create discrepancies between the choices we make and the feedback we get? Would we always

[1]Lund University Cognitive Science, Lund University, Kungshuset Lundagård, 222 22 Lund, Sweden. [2]Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: lars.hall@lucs.lu.se

R E P O R T S

be able to tell if an error were made? And if not, what would we think, and what would we say?

To examine these questions, we created a choice experiment that permitted us to surreptitiously manipulate the relationship between choice and outcome that our participants experienced. We showed picture pairs of female faces to 120 participants (70 female) and asked them to choose which face in each pair they found most attractive. On some trials, immediately after their choice, they were asked to verbally describe the reasons for choosing the way they did. Unknown to the participants, on certain trials, a double-card ploy was used to covertly exchange one face for the other (Fig. 1). Thus, on these trials, the outcome of the choice became the opposite of what they intended. Each subject completed a sequence of 15 face pairs, three of which were manipulated (M). The M face pairs always appeared at the same position in the sequence, and for each of these pairs, participants were asked to state the reasons behind their choice. Verbal reports were also solicited for three trials of non-manipulated (NM) pairs (11).

The experiment employed a 3-by-2, between-group factorial design, with deliberation time and similarity of the face pairs as factors. For time, three choice conditions were included: one with 2 s of deliberation time, one with 5 s, and one where participants could take as much time as they liked. Participants generally feel that they are able to form an opinion given 2 s of deliberation time (supporting online text). Nevertheless, the opportunity for participants to enjoy free deliberation time was included to provide an individual criterion of choice. For similarity, we created two sets of target faces, a high-similarity (HS) and a low-similarity (LS) set (fig. S1). Using an interval scale from 1 to 10, where 1 represents "very dissimilar" and 10 "very similar," the HS set had a mean similarity of 5.7 (SD = 2.1) and the LS set a mean similarity of 3.4 (SD = 2.0).

Detection rates for the manipulated pictures were measured both concurrently, during the experimental task, and retrospectively, through a post-experimental interview (11) (supporting online text). There was a very low level of concurrent detection. With a total of 354 M trials performed, only 46 (13%) were detected concurrently. Not even when participants were given free deliberation time and a set of LS faces to judge were more than 27% of all trials detected this way. There were no significant differences in detection rate between the 2-s and 5-s viewing time conditions, but there was a higher detection rate in the free compared to the fixed viewing time conditions [$t(118)$ = 2.17, P = < 0.05]. Across all conditions, there were no differences in detection rate between the HS and the LS sets (Fig. 2A). In addition, there were no significant sex or age differences in detection rate. Tallying all forms of detec-

tion across all groups revealed that no more than 26% of all M trials were exposed.

However, these figures are inflated even so. The moment a detection is made, the outlook of the participants changes: They become suspicious, and more resources are diverted to monitoring and control. To avoid such cascading detection effects, it is necessary to discard all trials after the first detection is made. Figure 2B shows detection rates with this correction in place. The overall detection rate was significantly lower [$t(118)$ = 3.21, P < 0.005], but none of our prior conclusions are affected by the use of this data set (the percentage of participants that detected the manipulation is shown in fig. S2).

Our experiment indicates that the relationship between intentions and outcomes may sometimes be far looser than what current theorizing has suggested (6, 9). The detection rate was not influenced by the similarity of the face pairs, indicating the robustness of the finding. The face pairs of the LS set bore very little resemblance to each other, and it is hard to imagine how a choice between them could be confused (fig. S1 and supporting online text). The overall detection rate was higher when participants were given free deliberation time.

This shows the importance of allowing individual criteria to govern choice, but it is not likely to indicate a simple subjective threshold. The great majority of the participants in the 2-s groups believed themselves to have had enough time to make a choice (as determined by post-test interviews), and there was no difference in the actual distribution of choices among the pairs from fixed to free deliberation time.

Next, we examined the relationship between choice and introspective report. One might suspect that the reports given for NM and M trials would differ in many ways. After all, the former reports stem from a situation common to everyday life (revealing the reasons behind a choice), whereas the latter reports stem from a truly anomalous one (revealing the reasons behind a choice one manifestly did not make).

We classified the verbal reports into a number of different categories that potentially could differentiate between NM and M reports. For all classifications, we used three independent blind raters, and interrater reliability was consistently high (supporting online text and table S1). We found no differences in the number of empty reports (when participants were unable to present any reasons at all) or in the

Fig. 1. A snapshot sequence of the choice procedure during a manipulation trial. (**A**) Participants are shown two pictures of female faces and asked to choose which one they find most attractive. Unknown to the participants, a second card depicting the opposite face is concealed behind the visible alternatives. (**B**) Participants indicate their choice by pointing at the face they prefer the most. (**C**) The experimenter flips down the pictures and slides the hidden picture over to the participants, covering the previously shown picture with the sleeve of his moving arm. (**D**) Participants pick up the picture and are immediately asked to explain why they chose the way they did.
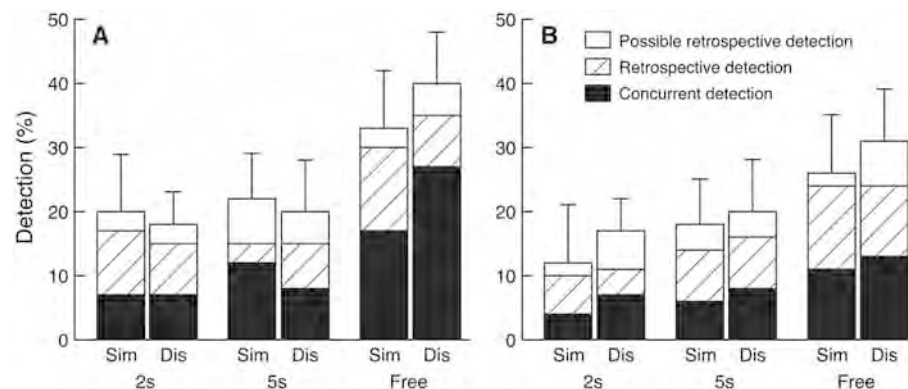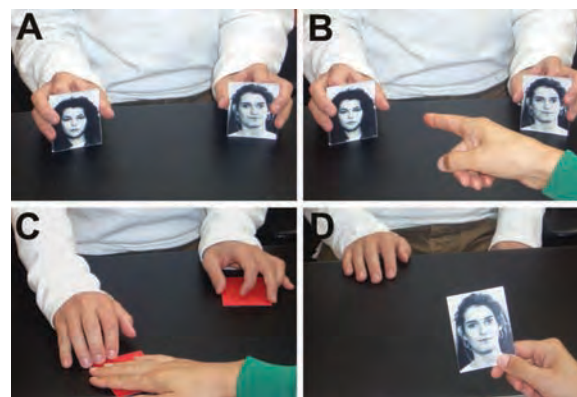
Fig. 2. Percent detection, divided into deliberation time and similarity, for (**A**) all trials and (**B**) trials corrected for prior detections. Sim, similar (HS); Dis, dissimilar (LS). Error bars indicate the standard deviation of the means.

www.sciencemag.org   SCIENCE   VOL 310   7 OCTOBER 2005   **117**

degree to which reports were phrased in present or past tense (which might indicate whether the report is made in response to the present face or the prior context of choice). Neither did the length of the statements, as measured by number of characters, differ between the two sets (NM = 33, SD = 45.4; M = 38, SD = 44.4), nor the amount of laughter present in the reports (with laughter being a potential marker of nervousness or distress). We found significantly more dynamic self-commentary in the M reports [$t(118) = 3.31$, $P < 0.005$]. In this type of commentary, participants come to reflect upon their own choice (typically by questioning their own prior motives). However, even in the M trials, such reports occurred infrequently (5% of the M reports).

We rated the reports along three dimensions: emotionality, specificity, and certainty (using a numeric scale from 1 to 5). Emotionality was defined as the level of emotional engagement in the report, specificity as the level of detail in the description, and certainty as the level of confidence in their choice the participants expressed. There were no differences between the verbal reports elicited from NM and M trials with respect to these three categories (fig. S3). Seemingly, the M reports were delivered with the same confidence as the NM ones, and with the same level of detail and emotionality. One possible explanation is that overall engagement in the task was low, and this created a floor effect for both NM and M reports. However, this is unlikely to be the case. All three measures were rated around the midline on our scale (emotionality = 3.5, SD = 0.9; specificity = 3.1, SD = 1.2; certainty = 3.3, SD = 1.1). Another possibility is that the lack of differentiation between NM and M reports is an indication that delivering an M report came naturally to most of the participants in our task. On a radical reading of this view, a suspicion would be cast even on the NM reports. Confabulation could be seen to be the norm and truthful reporting something that needs to be argued for.

To scrutinize these possibilities more closely, we conducted a final analysis of the M reports, adding a contextual dimension to the classification previously used. Figure 3 shows the percentage of M reports falling into eight different categories. The "specific confabulation" category contains reports that refer to features unique to the face participants ended up with in a manipulated trial. As these reports cannot possibly be about the original choice (i.e., "I chose her [the blond woman] because she had dark hair"), this would indeed be an indisputable case of "telling more than we can know" (12). Equally interesting is the "original choice" category. These are reports that must be about the original choice, because they are inconsistent with the face participants ended up with (i.e., "I chose her because she smiled [said about the solemn one]". Here, despite the imposing context of the manipulated choice, vestiges of the original intention are revealed in the M reports. Analogous to the earlier example of confabulation, this would be an unquestionable case of truthful report.

In summary, when evaluating facial attractiveness, participants may fail to notice a radical change to the outcome of their choice. As an extension of the well-known phenomenon of change blindness (13), we call this effect choice blindness (supporting online text). This finding can be used as an instrument to estimate the representational detail of the decisions that humans make (14). We do not doubt that humans can form very specific and detailed prior intentions, but as the phenomenon of choice blindness demonstrates, this is not something that should be taken for granted in everyday decision tasks. Although the current experiment warrants no conclusions about the mechanisms behind this effect, we hope it will lead to an increased scrutiny of the concept of intention itself. As a strongly counterintuitive finding, choice blindness warns of the dangers of aligning the technical concept of intention too closely with common sense (15, 16).

In addition, we have presented a method for studying the relationship between choice and introspection. Classic studies of social psychology have shown that telling discrepancies between choice and introspection can sometimes be discerned in group-level response patterns (12) but not for each of the individuals at hand. In the current experiment, using choice blindness as a wedge, we were able to "get between" the decisions of the participants and the outcomes with which they were presented.

Type	%		
Specific Conf.	13.3		She's radiant. I would rather have approached her at a bar than the other one. I like earrings! [M]
Detailed Conf.	17.3	She looks like an aunt of mine I think, and she seems nicer than the other one. [F]	
Emotional Conf.	9.3		Yes, well, [laughter] she looks very hot in this picture. [M]
Simple Conf.	10.8		Just a nice shape of the face, and the chin. [M]
Relational Conf.	21.3		I thought she had more personality, in a way. She was the most appealing to me. [F]
Uncertainty	11.6	Eh... I don't know. [F]	
Dynamic report	5.2		Oh, [short laughter] Why did I choose her? She looks very masculine! [M]
Original choice	11.2	Because she was smiling. [F]	

Fig. 3. Frequency distribution of the contents of the M reports aligned along a rough continuum from confabulatory to truthful report. Sample sentences (translated from Swedish) are drawn from the set of reports for the displayed face pair. Letters in brackets indicate whether the report was given by a male (M) or a female (F) participant. The specific confabulation (Conf.) category contains reports that refer to features unique to the faces participants ended up with in an M trial. The detailed and emotional confabulation categories contain reports that rank exceptionally high on detail and emotionality (>4.0 on a scale from 1 to 5). The simple and relational confabulation categories include reports where the generality of the face descriptions precluded us from conclusively associating them with either of the two faces (i.e., everybody has a nose, and a personality). The category of uncertainty contains reports dominated by uncertainty (<2 on a scale from 1 to 5). The dynamic reports are reports in which participants reflect upon their own choice, and the final category contains reports that refer to the original context of choice.

This allowed us to show, unequivocally, that normal participants may produce confabulatory reports when asked to describe the reasons behind their choices. More importantly, the current experiment contains a seed of systematicity for the study of choice and subjective report. The possibility of detailing the properties of confabulation that choice blindness affords could give researchers an increased foothold in the quest to understand the processes behind truthful report.

References and Notes
1. K. R. Ridderinkhof, W. P. M. van den Wildenberg, S. J. Segalowitz, C. S. Carter, *Brain Cogn.* **56**, 129 (2004).
2. K. R. Ridderinkhof, M. Ullsberger, E. A. Crone, S. Nieuwenhuis, *Science* **306**, 443 (2004).
3. M. Ullsperger, D. Y. Cramon, *Cortex* **40**, 593 (2004).
4. M. E. Walton, J. T. Devlin, M. F. S. Rushworth, *Nat. Neurosci.* **7**, 1259 (2004).
5. P. Haggard, S. Clark, *Conscious. Cogn.* **12**, 695 (2003).
6. R. Grush, *Behav. Brain Sci.* **27**, 377 (2004).
7. D. M. Wolpert, K. Doya, M. Kawato, *Philos. Trans. R. Soc. London Ser. B.* **358**, 593 (2003).
8. J. G. Kerns *et al.*, *Science* **303**, 1023 (2004).
9. R. Hester, C. Fassbender, H. Garavan, *Cereb. Cortex* **14**, 986 (2004).
10. H. C. Lau, R. D. Rogers, P. Haggard, R. E. Passingham, *Science* **303**, 1208 (2004).
11. Materials and methods are available as supporting material on *Science* Online.
12. R. E. Nisbett, T. D. Wilson, *Psychol. Rev.* **84**, 231 (1977).
13. R. A. Rensink, *Annu. Rev. Psychol.* **53**, 245 (2002).
14. D. J. Simons, R. A. Rensink, *Trends Cogn. Sci.* **9**, 16 (2005).
15. D. C. Dennett, *The Intentional Stance* (MIT Press, Cambridge, MA, 1987).
16. D. M. Wegner, *The Illusion of Conscious Will* (MIT Press, Cambridge, MA, 2003).
17. We thank D. de Léon, K. Holmqvist, P. Björne, P. Gärdenfors, T. Dickins, N. Humphrey, A. Marcel, Q. Rahman, E. Adams, N. Bolger, B. Cohen, and E. Phelps for useful comments and discussions. We also thank C. Balkenius for providing the illustrations for the article and P. Rosengren for invaluable advice concerning the use of card magic techniques. Supported by The New Society of Letters at Lund (P.J.), The Knowledge Foundation, The Erik Philip-Sörensen Foundation (L.H.), and the Swedish Research Council (S.S.).

Sexual Selection Can Resolve Sex-Linked Sexual Antagonism

Arianne Y. K. Albert* and Sarah P. Otto

Sexual selection is a potent evolutionary force. However, very few models have considered the evolution of female preferences for traits expressed in both sexes. Here we explore how female preferences coevolve with sexually antagonistic traits, which involve alleles that are beneficial to one sex but harmful to the other. We show that with a sexually antagonistic trait on the X chromosome (males XY, females XX), females evolve to prefer mates carrying alleles beneficial to daughters. In contrast, with a Z-linked trait (males ZZ, females ZW), females more often evolve mating preferences for mates carrying alleles beneficial to sons (that is, flashy displays).

Evolutionary biologists have long puzzled over how and why female preferences drive the evolution of exaggerated male traits. Generally, female preferences are thought to enhance a female's long-term fitness by increasing her offspring's fitness, either directly or through genetic associations between preference and trait loci (*1*, *2*). Nearly all models assume that females do not initially express the male display trait, or else they assume that the fitness effects are the same in males and females (*3*). However, traits subject to sexual selection will often be sexually antagonistic, for example, with "sexy" male traits benefiting males but reducing female fitness (*4*–*6*). If a trait increases male reproductive success at the cost of female viability, females must then choose between having attractive sons (and unfit daughters) and having ugly sons (but unencumbered daughters). As long as females can detect the genotypic differences among males at sexually antagonistic loci, we expect mating preferences to evolve as described by the models explored here.

Theory predicts that sexually antagonistic loci are more likely to remain polymorphic on the sex chromosomes (*4*, *5*). Furthermore, recent empirical work suggests that many sexually selected traits in animals are located on the X chromosome (*7*, *8*) and that most polymorphic sexually antagonistic traits are located on the X chromosome in *Drosophila* (*5*, *9*). There is also evidence to suggest that a female's mate choice may result in a tradeoff in the fitness between her daughters and sons (*10*). Several recent theoretical examinations of the evolution of female preferences have explored sex linkage of the trait and/or preference (*11*–*14*). However, these models, with the exception of Reeve and Pfennig's (*12*),

assume that sexually selected traits have male-limited expression and therefore no fitness consequences when in females, and none has addressed sexual antagonism. Here we address the question of how female preferences evolve for traits that have contrasting fitness effects in each sex.

With sexual antagonism, chromosomal location should strongly affect the evolution of female preferences. Simply put, an X-linked male trait is never passed on from an attractive father to his sons, whereas his daughters suffer the cost of carrying the display trait (*5*, *9*). Offspring in XY species therefore do not gain a fitness benefit from females preferring males with a more extreme X-linked display trait. In contrast, both males and females contribute a Z chromosome to sons in ZW species. Thus, females preferring a Z-linked display trait receive the fitness benefit of sexy sons, even though their daughters suffer a fitness cost (*5*, *9*). This cost is lessened by the fact that daughters inherit only one of their father's Z chromosomes. With autosomal inheritance, these asymmetries in inheritance are absent.

To verify the verbal argument laid out above, we present the results of two-locus models that follow the fate of a newly arisen preference allele p in a population that is at a polymorphic equilibrium at a trait locus. We assume that

Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada.

*To whom correspondence should be addressed. E-mail: albert@zoology.ubc.ca

Table 1. Male and female fitness components in male heterogametic (XY) and female heterogametic (ZW) species.

		X-linked trait			Z-linked trait	
		Male trait				
		T	*t*		*TT*	*Tt*	*tt*
Female preference	*PP*	1	$1 + a_{PP}$	*P*	1	$1 + da_P$	$1 + a_P$
	Pp	1	$1 + a_{Pp}$	*p*	1	$1 + da_p$	$1 + a_p$
	pp	1	$1 + a_{pp}$				
Male viability		$1 - s_y$	1		$1 - s_z$	$1 - hs_z$	1
		Female trait				
		TT	*Tt*	*tt*	*T*	*t*
Female viability		1	$1 - hs_x$	$1 - s_x$	1	$1 - s_w$
		Female preference				
		PP	*Pp*	*pp*	*P*	*p*
Female cost		$1 - |a_{PP}|k$	$1 - |a_{Pp}|k$	$1 - |a_{pp}|k$	$1 - |a_P|k$	$1 - |a_p|k$

Supporting Online Material for

Failure to Detect Mismatches Between Intention and Outcome in a Simple Decision Task

Petter Johansson, Lars Hall,* Sverker Sikström, Andreas Olsson

*To whom correspondence should be addressed. E-mail: lars.hall@lucs.lu.se

This PDF file includes:

Materials and Methods
SOM Text
Figs. S1 to S3
Table S1
References and Notes

Supporting Online Material

Material and Methods

Participants
One hundred and twenty participants (70 female) participated in the study (mean age ± SD, 26 ± 8.3). Participants were drawn from a mixed student and non-student population. As a cover story for the experiment participants were told that the experimenters were interested in choice and facial attractiveness. After the experiment participants were debriefed about the true nature of the design, and given the opportunity to voice any concerns. All participants then gave informed consent. Two participants were removed from the subsequent analysis because they were immediately able to discern how the card trick was performed (due to flawed presentations by the experimenter).

Experimental Procedure
Participants were shown pairs of grayscale pictures of female faces, and were given the evaluative task of choosing which face in each pair they found most attractive. In addition, on some trials, immediately after the choice, they were asked to verbally describe the reasons for choosing the way they did. Participants had been informed in advance that we would solicit verbal reports about their intentions during the experiment, but not the specific trials for which this was the case. Unknown to the participants, on certain trials, a double-card ploy was used to covertly exchange one face for the other. Thus, on these trials, the outcome of the choice became the opposite of what they intended.

The experiment employed a three by two factorial design, with deliberation time and similarity of the face-pairs as factors. 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, we created two sets of target faces, a high similarity (HS) and a low similarity set (LS). Using an interval scale from 1–10 where 1 represents "very dissimilar" and 10 "very similar", the HS set had a mean similarity of 5.7 (SD=2.08), and the LS a mean similarity of 3.4 (SD=2.00). The face pictures were collected from the The Psychological Image Collection at Stirling (PICS), online face database (http://pics.psych.stir.ac.uk/). We used pictures from the Nottingham and the Stirling collection, and 15 face-pairs were constructed on the basis of a rough matching of the photos (position of the head, background luminance, background color, attractiveness, etc.). After this, a group of independent raters (n=15) coded all pairs for similarity, and six pairs were selected for the HS and LS set.

Each participant completed a sequence of 15 face-pairs, three of which were manipulated. The manipulated face-pairs always appeared at the same position in the sequence (7, 10, 14), and in the same order. For all of these pairs participants were asked to state the reasons behind their choice. All reports were recorded and later transcribed. To provide a comparison class, verbal reports were also solicited for three trials of non-manipulated pairs. The non-manipulated (NM) and manipulated (M) pairs were counter-

balanced during the experiment (with the LS set serving as non-manipulated control in the HS-groups, and equally the other way around).

Using standard change blindness terminology, this task would be described as involving incidental change detection, one-shot stimulus presentation, and occlusion-contingent change (*1*). The period the hidden picture remained unseen on the table during the switch was approximately 2s from drop-down to pick-up (with some variations due to natural arm movements).

Detection rates for the manipulated pictures were measured both concurrently and retrospectively, with three graded levels of detection being used for our analysis. A trial was classified as concurrently detected if participants showed any signs of detection during the switch (such as explicitly reporting that the faces had been switched, or indicating that something went wrong with their choice). After the experiment all participants were asked a series of increasingly specific questions in a post-test interview to investigate whether they had any inkling that something had gone wrong ("What did you think about the experiment?", "Did you find anything odd with the experiment?" and "Did you notice anything strange with the stimuli presented in the experiment?"). Participants that revealed no signs of detection in this procedure were then presented with a hypothetical scenario describing an experiment in which the faces they choose between are surreptitiously switched (i.e. the very experiment they had just participated in), and asked whether they thought they would have noticed such a change. This question was included to determine the folk-psychological status of our design (i.e. whether it would be perceived as counter-intuitive or not). Finally, all participants were debriefed about the true nature of the design, and asked if they had noticed anything in the experiment resembling the switches that we had just described. If they answered "no" to this question, we concluded that they did not consciously notice any of the manipulations made during the experiment. All other participants were then given an opportunity to sort through their chosen pictures and indicate which faces they felt could have been manipulated. A trial was classified as retrospectively detected if participants picked out the corresponding manipulated picture in the set. If participants did so, but also indicated any number of false positives, those trials were classified in a category called possible retrospective detection. The inclusion of this category in the analysis was meant to compensate for the possibility of underreporting due to unknown social factors present in the interview.

Supporting Text

Detection Criteria
Taken together we believe the three categories of detection in our experiment gave the participants a fair chance to voice their concerns, and that they go a long way towards ensuring that no conscious detections were left out. In devising a cued-procedure (i.e. allowing participants to sort through their chosen faces) for the retrospective detection test, and the inclusion of participants that even named false positives in the possible retrospective detection category, we tried to err on the side of being too liberal about what to count (for example, if we had terminated our post-test interview after the initial question

about whether participants experienced anything odd during the experiment, only a single retrospective detection would have been registered).

However, when discussing detection criteria it is very difficult to remain neutral with respect to different theories of consciousness. For our concurrent detection criterion we relied on spontaneous verbal report by the participants (even if we did not demand an articulate response). But why should we give special status to verbal reports? According to a prominent tradition in the field of implicit learning we should always be looking for the most exhaustive measure of conscious processing (*2-4*), otherwise we might end up establishing false dissociations between differentially sensitive measures of the same conscious resource. This methodological principle has been dubbed the *sensitivity criterion* (*4*).

The customary way of adhering to the sensitivity criterion is to use concurrent forced-choice to measure conscious detection (*5*). Applied to the current experiment this method would probably have resulted in more instances of detected manipulations than the spontaneous reporting we relied on. However, as we see it, there is a substantial difference between being unaware of a specific influence in a natural context, and being similarly unaware of some stimuli, influence, or process under the most penetrating probe (i.e. what the sensitivity criterion prescribes). The experiment was meant to simulate a choice situation in which no prior evidence indicates that a high level of monitoring is needed, and it is only very rarely that natural conversations are accompanied by clever simultaneous forced choice questions and reaction time measures to exhaustively probe our conscious knowledge.

Of course, any attempt at an ecological explanation of decision making would have to accommodate both non-vigilant (relaxed, non-suspicious), as well as vigilant (guarded, suspicious) choice. Depending on whether the correction for prior detection is applied in our experiment it can be seen to occupy different positions along this dimension, with the uncorrected version situated further towards the suspicious pole. Had our experiment been framed as an explicit detection task, we have no doubt that most participants would have been able to spot the manipulations immediately.

Previous Studies

Before implementing our main experiment we ran a series of basic studies exploring the phenomenon of choice blindness. These studies add to the evidential base of the current experiment by demonstrating the effect in a different medium and with a different design, and with different types of stimuli.

First, we created an experiment in which participants had to choose which one of two abstract patterns presented on a computer screen they found most aesthetically appealing (the patterns were collected from various websites containing 'artistic' computer wallpaper for non-commercial use). Each trial began when the participants clicked on a left-aligned start-icon that made the two patterns appear on the right side of the screen. Participants were given 1500 ms to consider their choice, then an alerting sound was played, and they had to move the cursor to the preferred pattern. In addition, we required the cursor trajectory to the target pattern to pass through one of two small, color-coded, intermediate squares corresponding to either the upper or the lower pattern on the right. When the participants passed through one of these squares, the entire screen flashed in matching color

for 50 ms. Similarly to the current experiment, on some trials, a mismatch between choice and outcome was created. On a manipulated trial, the attention-grabbing properties of the midway square and the 50 ms screen flash were used to conceal the fact that the two choice alternatives switched places while the participants were moving the cursor across the screen. The full experiment consisted of 15 trials, three of which were manipulated. Twenty participants (12 female) were tested. In total, counting both concurrent and retrospective detections (and using data uncorrected for prior detections), 19% of the manipulated trials were detected.

In a subsequent experiment we used the same decision paradigm, but instead of abstract patterns we used female faces to choose between (as with the current experiment, the PICS online face database was used for the selection). In addition, immediately after their choice participants were asked to state their reasons for choosing the way they did. The experiment consisted of 30 trials, five of which were manipulated. Twenty-two participants (14 female) were tested, and the total detection rate was 32%. However, with five manipulated trials used rather than three, prior detection made a larger impact on the detection rate. Using corrected data, detection rate drops to 20% (this can also be seen in the fact that 9 out of 22 participants did not detect any of the five manipulations). Analysis of the verbal reports revealed similar patterns as in our main experiment, with no clear differentiation between the NM and M-reports.

Finally, we used the same setup as in the previous experiment, but with a set of male faces to choose between (again, the faces were collected from the PICS database). In addition, eye-tracking was used to verify that participants attended to the pictures both during the deliberation phase, and when giving their verbal reports. Eighteen participants (12 female) were tested, and total detection rate was 37% (29%, when corrected for prior detection). Analysis of the eye-tracking data revealed that participants attended to the pictures both before and after their choice. Again, analysis of the verbal reports revealed no differences between the NM and M-trials.

Throughout the whole series of studies, and in pilot controls, we conducted post-experiment interviews to determine the subjective confidence participants felt about their choices. While opinion about whether the task was difficult or not fluctuated somewhat, a great majority of the participants believed 1500 ms was enough time to make a proper choice.

Choice Blindness Blindness
When we claim that it is hard to believe how a choice between the face-pairs in our study could be confused, we are not simply asking our readers to inspect the pairs in Fig. S1 and form their own opinions. During the post-test interview in the experiment we requested all participants that had not yet voiced any suspicion to consider a hypothetical choice-manipulation extension of our experiment (see above, *experimental procedure*) and asked them if they believed they would have noticed such a change. The result shows that of the participants in our study that failed to notice any of the manipulations, 84% believed that they would have been able to do so (a result comparable to similar metacognitive probes in the change blindness literature (*6, 7*). Accordingly, many participants also showed considerable surprise, even disbelief at times, when we debriefed them about the true nature of the design. This effect of "choice blindness blindness" was also evident in our earlier

computer-based experiments, with roughly 87% percent of participants claiming that they would have noticed if the outcome of their choice had been manipulated in the hypothetical experiment we described.

Analysis of the Introspective Reports
Analysis of verbal reports often proceeds in several iterations, where the early rating results are used to distill a more distinct and consistent categorization (*8-10*). The contrastive analysis we employed to analyze potential differences between the NM and M-trials, were based on a two-stage classification of the verbal reports of our participants. As the NM-reports stem from a situation common to everyday life, while the M-reports are produced in response to a truly anomalous experimental probe, it would be natural to suspect that the two types of reports would differ in many ways. To investigate this, we identified four simple variables, based on 'surface' features of the reports (empty reports, laughter, the length of the reports, and the tense of the reports), and four promising psychological dimensions (emotionality, specificity, certainty, and dynamic self-reference). For all of these items common sense would suggest that the NM- and M-reports ought to be differentiated: participants in the M-trials ought to be more likely to say "I don't know", or "I have no idea", when asked to state the reasons behind a choice they did not make (*empty reports*); they ought to give shorter reports (*length of report*); they ought to produce more nervous laughter or giggle in response to the unfamiliarity of the situation (*laughter*); and they ought to make more references to past tense in their reports, talking about what they thought in relation to the original context of choice, rather than what they think about the picture they are seeing now (*tense of report*); participants in the M-trials ought also to show less emotional engagement, as the M-reports are given in response to the alternative they did not prefer (*emotionality*); they ought to make less specific and detailed reports, as no prior reasons have been formulated for the manipulated alternative (*specificity*); they ought to express less certainty about their choice (*certainty*), and they ought to reflect more about the current choice situation, and engage in more dynamic self-commentary, typically by questioning their own prior motives (*dynamic report*).

Independent raters first made untrained judgments for the classifications and dimensions we had identified (except length of report, which we calculated using the spreadsheet software). Each rater coded the whole set of reports. Three raters coded the four simple variables, and we used another three raters for the more complex scales. Next, we consulted with the group of raters, and used their input to sharpen our criteria and to calibrate our scales. Then a second group of (3+3) independent raters was given the same task. Before the new rating procedure each rater was provided with a training kit containing definitions and examples (available upon request from the authors). The approximate amount of training and instruction given to the raters ranged from fifteen minutes for the simple categories, to approximately 45 minutes for the psychological dimensions. This procedure resulted in good interrater agreement (see discussion below).

The final contextual analysis proceeded somewhat differently. Here, we were interested in investigating the *relation* between the content of the M-reports and the picture they were presented with at the time of the report. More specifically, raters were given the task of classifying whether the reports contained references to unique or distinguishing features of one of the two faces in each pair – i.e. whether the report was *about* a particular

face. As with the other categorizations, this task was first given to three independent raters, then calibrated, and then given to another three raters for a final classification. However, as we wanted the classification to be unquestionable, we only included instances of reports in the final analysis for which the raters had absolute agreement.

The introspective reports collected in our experiment are rich and varied, and it is important not only to search for differences between the NM- and M-reports, but also to provide a descriptive representation of the content of these reports. In Fig. 3 we plot the frequency of eight different categories for the M-reports, laid out in a rough continuum between confabulatory and truthful report. The figure is built around epistemic 'anchor points' at each end (i.e. the categories 'specific confabulation' and 'original choice', for which we can be certain that the reports are either confabulatory or truthful), and then reports are collated according to the degree to which they are likely candidates to be confabulations. For example, a report saying "[I chose her] because she has a nice face" is placed at the center of the continuum. A report of this kind contains no information that allows us to assign it to either of the two choice alternatives (i.e. everybody has a face; it is not a distinguishing feature). Also, it has no additional interesting properties, like a strong emotional component, or a high degree of specificity. There are good reasons to believe this report in fact *is* a confabulation (after all, it is produced in direct response to a face the participant manifestly did not choose), but the content of the report gives no further clues about whether this is the case. In contrast, a report that is highly emotional, like "I simply love this girl", represents a more severe mismatch between the actual choice and the manipulated outcome, and is placed closer to the confabulatory pole. On the other hand, a report that is devoid of any content, like "I don't know", or "I can't tell", is marked by uncertainty, and is therefore placed further towards the truthful pole.

As mentioned above, a great strength of our methodology is that it allows for us to detect categories of reports in the M-trials that undoubtedly refer to the manipulated picture ("specific confabulation") or the original context of choice ("original choice"). But currently there is no way to make these distinctions for the NM-reports, which preclude any comparisons between NM- and M-reports for these two categories. The categorization in Figure 3 is mutually exclusive, and weighted by proximity to the two poles. Reports were first placed in the two outmost categories, then in the category of dynamic report, then in detailed confabulation, then according to emotionality, then uncertainty, and finally the rest of the reports were divided into the simple and relational categories. As we see it, the resulting distribution gives a highly interesting impression of the contents of the M-reports, revealing the variable nature of, and the varying tendencies for, truthful and confabulatory report by our participants.

To measure interrater reliability (IRR) we used Pearson's product moment correlation as our main index. Table S1 shows the IRR levels for all variables and dimensions used in our analysis. The IRR is based on the average of the pair-wise Pearson product moment correlations between the three raters. Pearson's *r* is a well-established index that measures internal consistency and covariation between raters. As we were mainly interested in investigating potential differences between the two classes of reports (NM and M), a covariation index is appropriate to use. However, it should be noted that estimates of IRR may fluctuate between different measurements. In the words of (*9*).: "Despite all the effort that scholars, methodologists, and statisticians have devoted to developing and testing

indices of intercoder reliability, there is no consensus on a single 'best' index" (p. 593). As (*9*) contend, it is advisable to calculate IRR using more than one measure, and to demonstrate consistency across measures. Thus, although r is a commonly applied statistic for estimating the IRR, we have also chosen to include calculations based on Intra Class Correlation (ICC), and Krippendorff's Alpha (see Table S1). The ICC is a measure widely endorsed to estimate IRR when ratings from more than two judges are considered (*11, 12*). We based our ICC on a two-way ANOVA, treating both the targets (verbal reports) and the raters as the random factors. Because systematic differences among levels of ratings were considered relevant, a measure of absolute agreement was chosen. In the terminology proposed by Shrout and Fleiss (*13*), we computed a case 2 model with three raters ($ICC_{2,3}$). Krippendorff's Alpha is a chance corrected index of absolute agreement, which generally is considered to be a 'conservative' measurement of IRR (*9, 14*). As with the methods used to calculate IRR, there are no absolute standards about what constitutes acceptable levels of reliability (*9*), but as a result primarily intended for research purposes, our IRR levels must be considered high (*14-16*).

From Change Blindness to Choice Blindness
It has been known for a long time that human participants are inept at noticing changes in a visual scene when the transients accompanying that change no longer convey information about its location, a phenomenon that has been termed *change blindness* (*17*). During the last decade the phenomenon of change blindness has generated an extraordinary amount of interest among researchers interested in the workings of the human visual system (*1*), particularly with reference to the mechanisms of attention (*18*), and the nature of visual consciousness (*19*). But despite this, the full potential of change blindness as a tool for studying the human mind is far from realized. Why should change blindness only be used to study distinctly *visual* aspects of human cognition? (*1*) writes: "the study of change detection has evolved over many years, proceeding through phases that have emphasized different types of stimuli and different types of tasks. *All studies, however, rely on the same basic design*. An observer is initially shown a stimulus… a change of some kind is made to this stimulus… and the [visual detection] response of the observer is then measured (p. 251, our emphasis)". We were interested in the possibility of modifying this basic design to incorporate other non-perceptual elements of cognition. In particular, we wanted to investigate the relationship between intention, choice, and introspection. Our approach involves embedding different forms of change-manipulations in simple decision tasks and concurrently probe participants about the reasons for their choice. We see three main reasons for why this constitutes a novel and significant extension of the change blindness literature.

Firstly, choice blindness brings the conceptual tools of change blindness from the basic study of perception into a new domain of inquiry. Research on change blindness has occasionally contained elements of interaction (most notably, the real-person interactions in *20, 21*), and at least one task in which the actions of the participants have functional relevance has been investigated (*22*), but ours is the first study to incorporate meaningful decision making in an evaluative task. In change blindness experiments participants are usually more likely to notice changes when they concern features of particular relevance to the scene, or if they are of central interest to the participants, or if the participants are

particularly knowledgeable about them (*1, 23*). For choices it would almost seem to be a defining feature that they concern properties of high relevance and interest, or things we are very knowledgeable about. But in the current experiment, in the great majority of trials, our participants were blind to the mismatch between choice and outcome. While intending to choose X (a central-interest, non-peripheral, valenced stimuli), they failed to take notice when ending up with Y. This is a result that ought to be surprising even to the most seasoned change blindness researcher. On a more general level, we believe decision making to be domain with immediate intuitive appeal. There can be no doubt that we often care deeply about what we choose. The fact that we may be blind to the outcome of these choices is a finding that potentially could change our most intimate conceptions of ourselves as decision makers.

Secondly, choice blindness can be used to study introspection and preference change. Looking at the wider methodological aspects of our work, we believe choice blindness opens up exiting new opportunities for research. During the course of a normal day humans make countless choices: some slow and deliberate, some rapid and intuitive, some that carry only minor significance, and some that impact greatly on our lives. But for all the intimate familiarity we have with everyday decision making, it is very difficult to probe the representations underlying this process, or to determine what we can know about them from the 'inside', by reflection and introspection (*24-26*). The greatest barrier for scientific research in this domain is the nature of subjectivity. How can researchers ever corroborate the reports of the participants involved, when they have no means of challenging them? As philosophers have long noted, incorrigibility is a mark of the mental (*27*). Who are *they* to say what *my* reasons are? But as we have shown in the current analysis, choice blindness can be used to investigate the properties of introspective report. Beyond the exploratory work reported here, we envisage the collection and construction of large scale databases of reports given in relation to NM- and M-trials. By varying stimulus, personality and situational dimensions within the body of reports, powerful systematic comparisons between NM- and M-reports will become possible (both hypothesis-based and of a more data-driven nature). It is our belief that this will allow researchers to find patterns of reporting that will enable them to say something about the general properties of introspective reports, something no other current method is able to reveal. However, this is not the only methodological possibility afforded by the phenomenon of choice blindness. For example, by extending our basic design to incorporate repeated decisions in longer series of trials, choice blindness can be used to gain insight into the interplay between decision and feedback, choice and report, attitude and outcome. In this vein we have shown how feedback from M-trials can induce *preference change*, and how this bias of future choices relates to the introspective reports given in the experimental situation (*28*).

Thirdly, different mechanisms may underlie choice blindness and change blindness. Given that the current behavioral study was not designed to address the neuro-cognitive underpinnings of either choice or change blindness, it would be premature to offer any speculations whether they indeed are identical. However, as we see it, our experiment is perfectly positioned to bridge the disconnected research areas of choice/intentionality and change blindness, and to create some productive friction between the two. This can be seen clearly by a brief exposition of what intentional choice is supposed to entail. (*29*) write: "*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" (p. 80, our emphasis, see also *30-32*). But if this is the case, how can it be that the participants in our study often failed to detect the glaring discrepancy between the prior intention and the outcome of their choice? Matching this question with the most common explanations for change blindness offered in the literature does not seem to produce any satisfactory answers. In fact, in our view, given the almost complete lack of reference to mechanisms of decision making and intentionality in the change blindness literature, choice blindness would be an even more remarkable phenomenon if it turned out to be qualitatively identical to change blindness.

For example, the prevalence of choice blindness in our experiment might be due to a failure to sufficiently encode the choice alternatives during the deliberation phase (*33*). But from the perspective of a decision researcher it would amount to a strangely maladaptive decision process not to encode the features that are supposed to be the very basis of the choice, or the gross identity of the two alternatives (at the very least, this should hold for the condition with free viewing time, where the participants themselves set the criteria for when to terminate the deliberation). Another option is that the intentions simply are forgotten during the two second interval when the card is switched. But intentions are not supposed to be instantly forgotten. As (*29*) contend, they are supposed to be the guiding structures behind our actions (and phenomenologically speaking, this is what many people claim them to be), which makes this option equally unattractive to decision theorists. Similar things can be said for the other common explanations for change blindness: that initial representations might be disrupted or overwritten by the feedback (*34*), that change blindness results from a failure to compare pre- and post-change information (*35, 36*), or that explicit change detection is impossible because the representations are in a format inaccessible to consciousness (*37*). They are all viable candidates to explain choice blindness, but also more or less incompatible with popular theories of choice and intentionality. If our task can be seen as a good example of willed action, involving perfectly standard intentions and choices (and currently we can see no reason why this should not be the case), but the outcome of the experiment could be fully explained by the conceptual apparatus of change blindness research, then something would seem to be seriously amiss in current theories of decision making and cognitive control.

Supporting Figures

Fig S1. The face-pairs used for the manipulated trials in the experiment, with similarity scores displayed below each pair. The High-Similarity group is shown on the left, and the Low-Similarity group on the right.
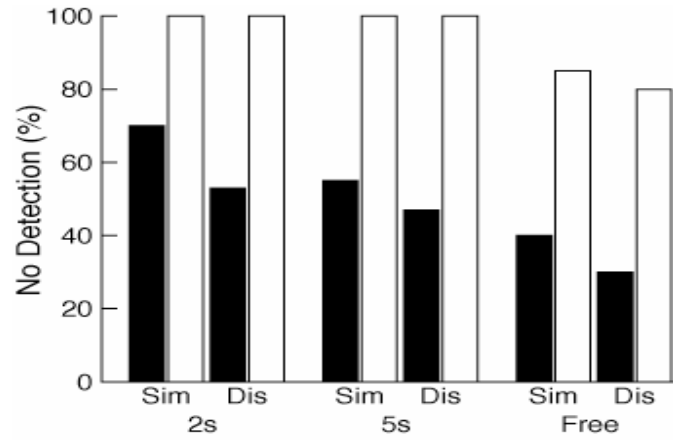
Fig S2. Percentage of subjects across the different conditions failing to detect all manipulations (black bars), and at least one manipulation (white bars).
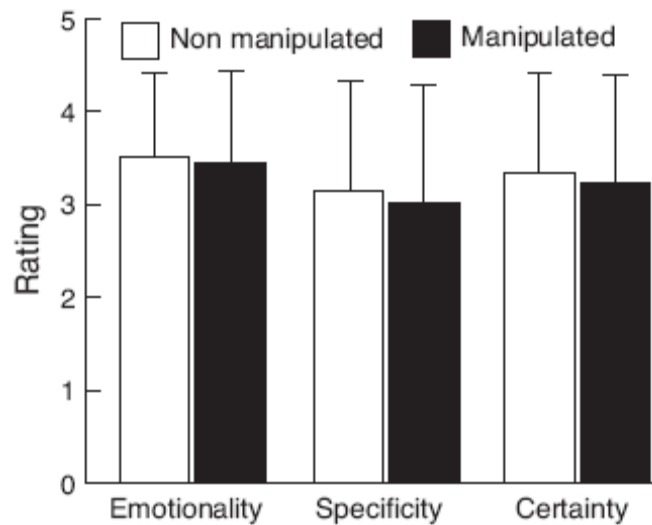
Fig S3. The content of the verbal reports rated along the dimensions of (A) emotionality, (B) specificity, and (C) certainty. As can be gleaned from the figure, no significant differences between the non-manipulated and the manipulated reports were found with respect to these three dimensions.

Supporting Tables

Variable	Pearson's r	ICC	Krippendorff
Laughter	0.95	0.98	0.95
Empty Reports	0.82	0.92	0.80
Tense	0.92	0.97	0.93
Emotionality	0.79	0.91	0.78
Specificity	0.88	0.96	0.88
Certainty	0.78	0.89	0.73
Dynamic Report	0.80	0.92	0.80
Specific Conf.	0.78	0.91	0.78
Original Choice	0.82	0.93	0.82

Table S1. Three alternative interrater reliability (IRR) measures for each variable used in our analysis. For each variable, the rating was performed by three independent raters. Listed in the left column are the average pair-wise Pearson product moment correlations between the three raters. The center column contains values for a case 2 model Intraclass Correlation (ICC) of absolute agreement, treating both the verbal reports and the raters as random factors. The right column contains values for Krippendorff's Alpha, a chance corrected index of absolute agreement, which is generally considered to be a conservative measurement of IRR. As can be seen in the figure, the IRR levels are uniformly high, with good consistency between measures.

Supporting References and Notes

1. R. A. Rensink, *Annu. Rev. Psychol.* **53**, 245 (2002).
2. P. M. Merikle, E. M. Reingold, *J. Exp. Psychol. Gen.* **127**, 304 (1998).
3. P. M. Merikle, M. Daneman, in *The New Cognitive Neuroscience*, M. S. Gazzaniga, Ed. (MIT Press, Cambridge, MA, ed. 2, 2000), pp. 1295-1303.
4. D. R. Shanks, M. F. St. John, **17**, 367 (1994).
5. P. F. Lovibond, D. R. Shanks, *J. Exp. Psychol. Anim. Behav. Processes* **28**, 3 (2002).
6. D. T. Levin, N. Momen, S. B. Drivdahl, D. J. Simons, *Visual Cogn.* **7**, 397 (2000).
7. B. J. Scholl, D. J. Simons, D. T. Levin, in *Thinking About Seeing: Visual Metacognition in Adults and Children*, D. T. Levin, Ed. (MIT Press, Cambridge, MA, 2004), pp. 145-164.
8. K. A. Neuendorf. *The content analysis guidebook.* (Sage, Thousand Oaks, CA, 2002)
9. M. Lombard, J. Snyder-Duch, C. C. Bracken. *Human Com. Res. 28,* (2002).
10. S. E. Stemler. *Practical Asses., Res. & Eval.* **9**, 4 (2004)
11. K. O. McGraw, S. P. Wong. *Psych. Methods*. **1**, 4. (1996).
12. J. Uebersax. *Statistical methods for rater agreement*. (Retrieved July 20, 2005, from http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm)
13. P. E. Shrout, J. L. Fleiss. *Psychological Bulletin* **86**, 420 (1979)
14. K. Krippendorf. *Human Commun. Res.* **30**, (2004).
15. K. Krippendorf. *Content Analysis: an Introduction to Its Methodology* (Sage, Thousand Oaks, CA, 2003).
16. J. L. Fleiss. *Statistical methods for rates and proportions* (2nd ed.) (John Wiley and Sons, NY, 1981).
17. D. J. Simons, D. T. Levin, *Trends Cogn. Sci.* **1**, 261 (1997).
18. P. U. Tse, D. L. Sheinberg, N. K. Logothetis, *Psychol. Sci.* **14**, 91 (2003).
19. A. Noe, *J. Conscious Studies* **9**, 5 (2002).
20. D. J. Simons, D. T. Levin. Psychon. *Bull. & Rev*. **5**. (1998).
21. D. T. Levin, D. J. Simons, B. Angleone, C. F. Chabris. *Brit. J. Psychol*. **93**, 289 (2002).
22. J. Triesch, D. Ballard, M. Hayhoe, B. Sullivan. *J. of. Vision*. **3** (2003).
23. S. Werner, B. Thies. *Visual. Cogn*. **7**. (2000).
24. R. E. Nisbett, T. D. Wilson, *Psychol. Rev.* **84**, 231 (1977).
25. T. D. Wilson, E. Dunn, *Annu. Rev. Psychol.* **55***,* 493 (2004).
26. A. I. Jack, A. Roepstorff, *J. Conscious. Studies* **11**, 7 (2004).
27. R. Rorty, *J. Philos* **67**, 399 (1970).
28. B. Tärning, L. Hall, P. Johansson, S. Sikström, A. Olsson. *Unpublished Manuscript*. (2005).
29. A. Sigiru et al. *Nat. Neurosci*. **4**. (2004).
30. M. Ullsperger, D. Y. Cramon, *Cortex* **40**, 593 (2004).
31. K. R. Ridderinkhof, W. P. M. van den Wildenberg, S. J. Segalowitz, C. S. Carter, *Brain Cogn.* **56**, 129 (2004).
32. P. Haggard. *TRENDS in Cog. Sci*. **9**, 6 (2005).
33. J. K. O'Regan, A. Noe. *Behavioral and Brain Sci*. **24**. (2001).
34. M. R. Beck, D. T. Levin. *Percept. & Psychophys.* **65** (2003).
35. S. R. Mitroff, D. J. Simons, D. T. Levin. *Percept. & Psychophys.* **66**, 8 (2004).

36. A. Hollingworth. *J. Exp. Psychol. Hum. Perc. & Perf.* **29** (2003).
37. D. J. Simons, M. Silverman, in. *The Visual Neurosciences,* L. M. Chalupa & J. S. Werner, Eds. (MIT Press, Cambridge, MA, 2004).