Synthese (2021) 198:5437–5454
https://doi.org/10.1007/s11229-019-02414-3

KNOWING THE UNKNOWN

Is choice blindness a case of self-ignorance?

Lisa Bortolotti[1] · **Ema Sullivan-Bissett**[1]

Received: 11 January 2019 / Accepted: 25 September 2019 / Published online: 28 September 2019

Abstract
When subject to the choice-blindness effect, an agent gives reasons for making choice B, moments after making the alternative choice A. Choice blindness has been studied in a variety of contexts, from consumer choice and aesthetic judgement to moral and political attitudes. The pervasiveness and robustness of the effect is regarded as powerful evidence of self-ignorance. Here we compare two interpretations of choice blindness. On the *choice error* interpretation, when the agent gives reasons she is in fact wrong about what her choice is. On the *choice change* interpretation, when the agent gives reasons she is right about what her choice is, but she does not realise that her choice has changed. In this paper, we spell out the implications of the two interpretations of the choice-blindness effect for self-ignorance claims and offer some reasons to prefer choice change to choice error.

Keywords Self-knowledge · Choice blindness · Decision making · Consumer choices · Moral attitudes · Political attitudes · Ignorance

1 Choice errors, choice changes, and confabulation

The choice-blindness effect has been studied in a variety of contexts, from consumer choice and aesthetic judgement to moral and political attitudes, by Petter Johansson, Lars Hall, and their colleagues at the Choice Blindness Lab (e.g., Hall and Johansson 2009; Hall et al. 2010, 2012, 2013; Johansson et al. 2006, 2008; Strandberg et al. 2018). The most common way to characterise the surprising results of the application of the choice-blindness paradigm is to say that, due to experimental manipulation, an agent can sincerely provide articulate and convincing reasons for a choice she did not make, and this is particularly striking because the choice defended with reasons is in tension with the choice the agent actually made just moments earlier.

✉ Lisa Bortolotti
l.bortolotti@bham.ac.uk

Ema Sullivan-Bissett
E.L.Sullivan-Bissett@bham.ac.uk

[1] Philosophy Department, University of Birmingham, Edgbaston B15 2TT, UK

Choice blindness is often described as a form of confabulation. When people confabulate, they tell a story that they believe to be correct, for instance, a story about why they made a certain choice, but the story is not grounded in the evidence (Turner and Coltheart 2010; Hirsten 2005; Bortolotti 2018). That is because people are not aware of some of the causal factors responsible for their choice. Clearly, there are important similarities between confabulation and choice blindness, and both phenomena have been discussed in the philosophical literature to argue for the limitations of self-knowledge claims.

However, in this paper we do not address confabulation head-on (we do so elsewhere, such as in Bortolotti 2018 and Sullivan-Bissett 2015). The main reason is that we believe that the phenomena of confabulation and choice blindness are significantly different (as we explain in Sect. 3). In *confabulation* Anna chooses chocolate ice-cream because her sister did, but when she is asked about her choice, she explains it by appealing to different reasons, e.g., that chocolate is her favourite ice-cream flavour. There is no doubt that Anna is aware of what her choice was, but what she gets wrong is the causal process leading to her choice. Anna confabulates reasons for her choice.

In *choice blindness* Anna is asked by her mum whether she wants vanilla ice-cream or chocolate ice-cream and she says: "Chocolate". Her mum mishears her and replies: "Vanilla is an excellent choice, my dear. Why did you choose it?" Anna answers that vanilla is a more delicate flavour than chocolate.[1] Now, it is not clear whether Anna knows what her choice was, because she seems to consent to her mum's incorrect choice attribution and, further, provides reasons for a choice that is different from the choice she explicitly made.

In the second scenario, to say that Anna *confabulates reasons for her choice* does not sound right. Rather, there are two ways in which we can describe what happens to Anna.[2] She either attributes to herself (and gives reasons for) a choice she did not make, ignoring what her 'real' choice was (choice error); or she makes two different choices in quick succession, and she only gives reasons for the latter (choice change). Not only is the challenge to self-knowledge different in the two scenarios—as confabulation does not threaten the correctness of one's choice attribution but simply the groundedness of the explanation provided for it—but it is also unclear in what respect choice blindness involves *confabulating reasons*.

We may wonder whether it matters if we are choice-blind. It is difficult to overestimate the value of choice in our personal and public lives. To a large extent, we identify ourselves (and other people identify us) with our choices, from the choice of an occupation to the choice of a life partner, from the choice of where to live, to the

[1] It is possible that in some real-life examples similar to the one described, a child confirms what her mother says not to contradict her or disappoint her—as may happen when we do what is expected of us for reasons related to social conformity or social desirability. For the example to be an example of choice blindness we need to assume that Anna offers what she regards as a genuine explanation for her choice when she explains to her mother why she chose vanilla.

[2] We do not rule out that alternative interpretations could be offered for the scenario. The two interpretations we introduce here and elaborate in the rest of the paper reflect the views currently discussed in the choice blindness literature.

choice of which party to vote for at the next political election. It is presumed that the choices we make reveal important features of our character and worldview. For all sorts of reasons, we are not always in a position to turn our preferences into choices, but the assumption is that, when we have the opportunity to do so, we manifest who we are in the choices we make. And other people respond to us—with admiration or condemnation—largely on the basis of the preferences that are reflected in our choices.

The role of choice in our personal and public lives would be greatly undermined if we were shown to attribute to ourselves choices we never made and defend them with reasons (choice error). It would also be unsettling to realise that our preferences can change dramatically in the space of a few moments as a result of manipulation, causing inconsistency between past and future choices (choice change). The discovery that our behaviour can be easily manipulated is unwelcome in either case, but in this paper we ask what exactly we are manipulated into doing, and whether we exhibit self-ignorance when we are choice blind.[3]

In choice error, we are manipulated into misattributing choices to ourselves, and in choice change we are manipulated into reversing our choices. If we were mistaken about what our choices are, then other people would be justified in questioning whether choices really are a clue to the kind of people we are. If we were ignorant about our choices, this would show that choices are not (always) reliable surface manifestations of our 'deeper selves'. If, however, we were manipulated into reversing our choices in the space of minutes, then it is not our claim to self-knowledge as such that would be threatened, but our stability and coherence as agents. We would not fail to realise what our choices are, but we would still fail to realise that our choices changed. Measures to prevent manipulation would need to be significantly different in the two cases: in choice error, we would need to avoid failures of self-knowledge by enhancing our capacity to recognise and remember which choices we made and keep track of the preferences that justify those choices; in choice change, we would need to avoid behavioural inconsistencies by resisting inaccurate third-party attributions and enhancing our stability and coherence.

In this paper, we make a case for the choice change interpretation of at least some of the experimental results on choice blindness. We are not arguing that choice change is *always* the best interpretation of the experimental results of the choice-blindness studies. However, we establish that choice change is a serious contender and provide some reasons to prefer this interpretation to the dominant interpretation, choice error, in at least some cases. We hope this will prompt future empirical work in this area to shed some light on how common the phenomenon of choice change might be and on how it can be distinguished from choice error. In general, our view

[3] The term 'choice blind' is a little unfortunate given that we are interested in the correct interpretation of the results from studies using this term (on the choice change interpretation, there's a sense in which we are *not* blind to our choices after all). Throughout the paper we will follow the literature in using the term 'choice blind' but note here that we intend it to be neutral between the choice error and the choice change interpretation. The term as used by us merely picks out the phenomenon identified by results.

is that the experimental results collected under the umbrella of choice blindness may well resist neat classification under a single interpretation.

In Sect. 2, we offer an example of what choice blindness may look like out of the lab, showing how the phenomenon, just like confabulation, has been taken to challenge the intuitive connection between defending a choice with reasons and having knowledge of that choice.

In Sect. 3, we briefly discuss how earlier studies on consumer choice and attitude shifts relate to the choice-blindness studies. Studies on choice blindness were devised as a refinement and development of the previous work on introspection effects, although the phenomenon they target is not the mere fact that we tend to confabulate reasons for our choices.

In Sect. 4, we focus on choice blindness for moral and political attitudes, fleshing out in more detail the two interpretations of choice blindness; choice error and choice change. We suggest that the studies in moral and political attitudes lend themselves particularly well to the choice change interpretation.

In Sect. 5, we discuss some reasons in favour of the choice change interpretation and ask whether it is in tension with one of the findings of the choice-blindness studies, that reasons cement preferences. We conclude by reflecting on the implications of the choice change interpretation for claims about self-ignorance.

2 Choices misattributed or changed?

In some of the philosophical literature on self-knowledge (e.g., Moran 2001), an agent's capacity to give reasons for a belief or choice is claimed to suggest that the agent *authors* that belief or choice, in the sense that she has a special, first-personal way to know what her belief or choice is, via the reasons she has for endorsing it as her own. The authorship model of self-knowledge has been challenged on the basis of the results of psychological studies on introspection effects, with philosophers arguing that we come to know what our minds are largely in the same way as a third person would, by interpreting our behaviour which is a fallible exercise, prone to biases (Carruthers 2005); and that in some circumstances a third person is in a better position to know what really goes on in our minds than ourselves (Lawlor 2003).[4]

In choice blindness (as in some of the other empirical challenges to the power of introspection and first-person authority) the connection between reason giving and self-knowledge is an explicit target. When we endorse with reasons a choice we made, we ascribe to ourselves preferences that justify that choice. If we endorse with reasons a choice we never made ascribing to ourselves preferences we do not have, then the role of reason giving in securing authorship and privileged access to our choices is under threat. Let's see how this works with an example.

Suppose that Helena wants to buy a new dress for her friend's cocktail party. In the shop, she tries on both a blue dress and a grey dress and chooses the blue dress because it is brighter. We can say that the reason Helena chose the blue dress is

[4] For a more detailed account of this exchange please see Bortolotti (2009) and Bortolotti (2018).

because she has a preference for wearing a brighter dress at her friend's party. Now suppose that at the till the shop assistant makes a mistake and places the grey dress in Helena's shopping bag instead of the blue dress. When Helena gets home, she opens the bag and finds the grey dress in it. She comes to believe that she chose the grey dress and she explains her alleged choice to her flatmate by saying that the grey dress is more elegant than the blue dress she tried on in the shop. We would normally believe that the fact that Helena can give reasons for her choice—reporting with honesty and conviction why she preferred one dress rather than the other—is evidence for the claim that she is fully aware of what her choice was and she authors it, ascribing a preference to herself that guided her action. But in this scenario the fact that Helena can give reasons for choosing the grey dress does not seem, after all, to provide evidence for the claim that Helena is aware of what her choice was and that she is the author of that choice.

Helena is choice-blind. The standard interpretation of the choice-blindness studies is that, due to manipulation, agents are likely to (1) misidentify what their choices are and (2) attribute to themselves preferences they do not have but that justify the choices they mistakenly identify as theirs. Seen in this light, the choice-blindness phenomenon involves two errors that can be characterised as misattributions: Helen cannot tell either what choice she made in the shop or what preferences she has about the sort of dress she should wear at her friend's cocktail party. In this interpretation, choice-blindness studies threaten the idea that, by giving articulate and convincing reasons for choosing A over B, people come to know about their choices in a first-personal (distinctive and authoritative) way.

There is an alternative interpretation of the experimental results of the choice-blindness studies which deserves more attention than it has received so far. It says that, when agents give articulate reasons for making a choice, they may fail to track their earlier choices but they do not misattribute choices to themselves, and they correctly report their current preferences. Rather, their choice changed as a result of the experimental manipulation.[5] The prediction is that, if they were asked again to choose between the same items they were shown before, participants would commit to a different choice from the choice they originally made (this prediction is confirmed in one of the choice blindness studies, Hall and Johansson 2008, as we shall discuss later). After giving reasons for allegedly choosing the grey dress, if Helena were back in the shop facing the blue and the grey dress again, she would choose the grey dress.

So, agents correctly report their preferences at the time of offering reasons, and by offering reasons for choice B they commit to a choice that is different from the one they made earlier, choice A—the fact that they offer reasons for choice B may also make it more likely that they will stick to that choice should they be asked to make the same or a similar choice in the future. Thus, although choice B is not a choice they ever explicitly made, it can affect their future decision making. The idea is that the choices change as a result of what agents believe they chose (Lopes 2014:

[5] This interpretation of Johansson et al. (2005) study on choosing faces is given by Lopes (2014: pp. 29–30).

p. 30). In this second interpretation, agents are still blind to something: they are not blind to what their choices are but to the fact that their choices changed and to the reason why they changed.

Thus, in either interpretation subjects exhibit *ignorance*: in choice error, they are ignorant about their choices; in choice change, they are ignorant about their change of choices and about the factors contributing to their choice changing. That is because agents do not realise that a manipulation took place, and that the choices they defended with reasons are an effect of their beliefs about what they chose, which in turn are a result of such manipulation.

Independent of the interpretation we adopt, the phenomenon is puzzling and reveals a vulnerability to manipulation whose extent we might have not imagined without appreciating the results of the choice-blindness studies. However, we will argue that in the choice change interpretation the agents' behaviour is no longer an obvious threat to self-knowledge intended as correct self-attribution. The agent in the manipulated condition makes choice A first, but—when B is presented to her as her own choice—she endorses B by providing reasons for choosing B. The agent prefers B to A at the time of justifying choice B, but she does not realise that the choice she is defending is not the choice she originally made, and that her choice has changed as a result of the experimenters manipulating her beliefs about her choices.

Given that either account of choice blindness raises serious concerns from an epistemic point of view, adopting the latter interpretation does not lead to an apology of human agency or a rationalisation of agents' behaviour. On the choice change interpretation, it is hard to avoid concluding that our choices are fickle and bound to change for no good reason.

3 Introspection effects and choice blindness

There is copious psychological research disclosing the ills of reason giving. In some contexts, when we give reasons, we offer ill-grounded explanations for questionable consumer choices (Nisbett and Wilson 1977); or we shift our attitudes towards our romantic relationships (Wilson and Kraft 1993), endorsing attitudes that are not predictive of our future behaviour.

3.1 Consumer choice

In the now classic study by Richard Nisbett and Timothy Wilson (1977), participants are asked to explain their consumer choices (either a choice between identical pairs of socks or a choice between slightly different nightgowns). When asked why they made their choice, participants answer by appealing to (what they took to be) features of the items, such as the socks having a better colour or the nightgown being softer. However, the choice was influenced by the position of the items: the chosen item was typically on the right-hand side of the person making the choice. According to Nisbett and Wilson, people are not aware of some of the cognitive processes underlying their choices (priming effects) and

come up with reasons for those choices based on plausibility considerations. In the study, if the reasons given are intended as an *explanation* for the choices, participants offer *ill-grounded* explanations. The explanations are ill-grounded because the facts to be explained did not occur for the reasons identified by the participants: one can safely infer from the experimental design that people did not choose the items for colour or texture but for their relative position.

In the philosophical literature, the results of the Nisbett and Wilson study and other studies on introspection effects have been conceived as a threat to self-knowledge, as is evident in the work by Peter Carruthers (2005), Robin Scaife (2014), and Krista Lawlor (2003). Yet, the studies do not show that participants misattribute choices to themselves. Rather, they show that participants are not able to identify the causal factors leading them to make a certain choice. Whether the evidence undermines self-knowledge claims depends on what we take self-knowledge to be. If the correct self-attribution of choices is sufficient for self-knowledge, then the participants' self-knowledge is not compromised. If self-knowledge requires not only correct self-attribution of choices, but also knowledge of how one's choices came about, then participants' self-knowledge is under threat (Bortolotti 2009, 2018).

In one of the choice-blindness studies on consumer choice (Hall et al. 2010), people were asked to sample two varieties of jam and choose their favourite. Moments later, participants were asked to sample (what was presented as) their preferred jam again and explain why they chose it. Experimenters manipulated the second sampling so that for their second tasting people were offered not the jam they just chose, but the one that they just rejected. Most people did not detect the switch and provided convincing reasons for the choice that was presented to them as theirs, even when the flavour of the two jams was very different, such as cinnamon-apple and grapefruit.

Despite the explicit intent of the choice-blindness paradigm to extend the results of the classic Nisbett and Wilson study, the phenomena involved are significantly different. In the choice-blindness experiment people are asked to explain a manipulated choice, say, for cinnamon-apple jam, when they originally chose grapefruit jam. When the agent gives reasons for choosing the cinnamon-apple jam, she is either (1) ascribing to herself a preference that she does not have, misidentifying her choice (choice error); or (2) ascribing to herself genuine preferences that she acquired as a result of the manipulation and signalling (choice change). The main difference between the two interpretations lies in whether the agent 'changed her mind' in the course of the experiment.

In the original introspection-effect studies, participants were asked to give reasons for the only choice that they had explicitly made. Participants neither misattributed their choices nor changed them—the focus was on whether the explanation of their choices reflected the factors determining their choices. As the choice-blindness paradigm raises questions about whether we misattribute choices to ourselves or whether our choices change, its focus is different from that of the original Nisbett and Wilson study.

3.2 Dual attitudes

The phenomenon of dual attitudes has been observed in several contexts, but here we shall focus on the attitudes people have towards their romantic relationships (Seligman et al. 1980; Wilson and Kraft 1993; Wilson et al. 1989; Ivan 2013). In one study (Wilson and Kraft 1993), participants were asked to assess the relationships with the people they were dating. Then, they were divided into two groups, and at the end they were asked for their attitudes once more, and for a prediction about the future of their relationships. In group 1, the intermediate task was to list reasons for the success or failure of the relationship. In group 2, participants were given a different, unrelated task.

Results showed that those participants who were asked for reasons for their relationship appraisals experienced an *attitude shift*. The reasons the participants offered did not support the attitudes they initially reported, and when participants were asked to assess the relationship again, their assessment differed from the one they initially gave. The experimenters' conclusion was that thinking about reasons brought about new attitudes that were significantly different from those initially reported. One interesting question is whether the new attitudes were likely to be reflected in the participants' long-term behaviour. In a study which included a follow-up interview, Wilson et al. (1984) found that participants who evaluated their relationship and made a prediction about it without being asked for reasons made more successful predictions than the participants who were asked for reasons.

Why do people change their minds about the success and quality of their relationships when they are asked for reasons for their relationship appraisals? One hypothesis is that, when people are asked to assess their relationship at the start of the experiment, they express how they feel *at the time*. But when they provide reasons for their attitudes, they mention facts about their relationships that are readily available to them and easy to report and share. So, people attend to different aspects of their relationships in the course of the experiment. As a result of the reason-giving task, it is possible that new attitudes are formed, and these may not be compatible with the originally reported attitudes. People may find themselves with conflicting attitudes, for instance predicting that the relationship would last, when the originally reported attitude towards the relationship was negative.

The results from such experiments allow for two interpretations. Which interpretation is preferred depends on whether we think that the attitude reported as a result of the reason-giving exercise is 'real', that is, whether it is an attitude the person genuinely has towards her relationship. If the attitude defended with reasons is not real, then the phenomenon demonstrates an obvious failure of self-attribution: people are mistaken about what their attitudes are (*attitude error*). If the attitude defended with reasons is genuine, the phenomenon demonstrates the instability of people's attitudes, showing that such attitudes can shift quite dramatically in a little amount of time (*attitude change*) and in virtue of how they are elicited. These two interpretations parallel the choice error and choice change interpretations of the experimental results from choice-blindness studies.

If only the initial reports are reports of people's 'real' attitudes, the reason-giving exercise is exposed as misleading. If the latter reports are reports of genuine but

newly acquired attitudes, then people do not make attribution errors as such, but their attitudes are shown to be unstable and vulnerable to manipulation. As in the consumer choice study, participants are not aware of some of the factors that caused them to adopt new attitudes (including the evidence manipulation that is part of the experimental setting). In attitude error, people do not know what their attitudes are. In attitude change, people are not mistaken about what their attitudes are. However, they are unaware that their attitudes changed and ignore what caused the change.

If we see choice blindness as a natural extension of the dual-attitude cases, the phenomenon could be more aptly called '*dual choices*'. Consider the following study. Participants were asked to choose between two pictures of individual faces they did not know, on the basis of the attractiveness of the faces (Johansson et al. 2005). Then participants were asked to give reasons for their choice, but the face presented to them as their choice was not the face they previously chose but the one they rejected. Most people failed to notice that their choice had been reversed. They went on to offer reasons for choosing the face that was not their original choice (Johansson et al. 2008).[6]

As in the case of dual attitudes, depending on whether we think the preferences the agent ascribes to herself when she gives reasons are 'real' preferences, the phenomenon points to either an obvious failure of self-knowledge (choice error), or a demonstration of the instability of choices (choice change). In the former case, agents make a mistake about what their choices are. In the latter case, agents change their mind in the course of the experiment about what their choice is. Which interpretation is most plausible?

In the case of the experiment on faces, Johansson and colleagues embrace a choice error interpretation, suggesting that participants are (literally) *blind to their choices*. However, when they present the results of other choice-blindness experiments, such as those concerning moral and political preferences (which we will

[6] The choice-blindness experiment with faces featured 120 participants (70 female, 50 male) who were presented with pairs of photographs of female faces. They were asked to select which they found the most attractive. On some trials, having made the choice, participants were immediately asked for reasons for their choice. This was the non-manipulated set-up. In the manipulated set-up, a double-card ploy was used allowing the experimenter to switch one face for another, which meant that the face presented to the participant as the selected face was not the selected face but the one that had just been rejected a few seconds earlier. Each participant completed a sequence of fifteen pairs of faces, and three of those fifteen were manipulated such that the opposite face was presented as their choice. The switches occurred at the same position in the sequence for every participant, and participants were always asked to state the reasons for their choices. Three time-conditions were included, one with two seconds deliberation time, one with five seconds, and one in which participants were free to take as much time as they wanted. Two similarity of faces conditions were included, one high and one low. Detections were classified as *concurrent* if they occurred during the task, and as *retrospective* if they occurred during a post-experimental interview (Johansson et al. 2008: p. 117). The results were as follows: of 354 manipulated trials, 46 (13%) were concurrently detected. When participants were given free deliberation time and low similarity faces (and this is where one might expect high detection), no more than 27% of trials were detected. Johansson and colleagues found no significant differences in the detection rate across the two second and five second time conditions, though they found a higher detection rate in the free time condition as compared with the fixed time conditions. They found no differences in detection rate between the high similarity and low similarity conditions, nor any significant differences in detection rate depending on sex or age (Johansson et al. 2008: pp. 117–18).

discuss in the next section), they talk about preferences being *reversed* and choice blindness showing that we are not set in our ways and can change our minds, which suggests a choice change interpretation.

Some have argued that choice change is the best interpretation for the experiment on faces too. Dominic Lopes claims that, in the manipulation condition of the experiment, participants give reasons for what has become their new choice (Lopes 2014). For Lopes, it is no surprise that preferences are changeable and amenable to manipulation. The newly ascribed preferences are real and are newly formed as a result of the participants' beliefs about the choices they made (Lopes 2014: pp. 29–30). Given that by and large participants do not detect the manipulation, they come to believe that they chose the face that is presented to them. Participants do not realise that their preference shifted. In the manipulation condition, reasons are often given in the past tense: "I *thought* she had more personality in a way", "I *chose* her because she smiled", or "I *chose* her because she had dark hair" (Johansson et al. 2005, p. 118). That strongly suggests that people believe that the choice they defend with reasons was their choice all along, failing to appreciate the role of the manipulation in determining their change of choice. In Lopes' view, the participants' belief that the choice presented to them was the one they had previously made *determines* the choice they ascribe to themselves when they give reasons.

> Perhaps subjects in the manipulation condition changed their preference as a result of their choice. Since they did not notice the manipulation, they believed that they chose the displayed face, and that fact determines their preference. On this hypothesis, […] subjects' reasons do not accord with their initial preference as revealed by their initial choice, but they do accord with their eventual preference as determined by what they took to be their choice. (Lopes 2014, pp. 29–30)

One important difference between the second attitude in the dual-attitude studies and the second choice in the choice-blindness cases is that the latter seems to be predictive of future behaviour, at least to an extent. When asked again to choose between the same faces at a later time (Johansson et al. 2008), those in the manipulated condition chose the face they provided reasons for choosing, and not the face they had originally chosen. This suggests that *reasons cement preferences* (Lopes 2014, p. 30). However, reasons *follow blindly along* preferences *instead of regulating them* (Lopes 2014, p. 30), pace Moran, and this is where the overlap is between the phenomenon of confabulation and that of choice blindness. A similar idea is argued for in Dan Ariely's influential research, which is captured by the slogan: *Action determines, not just reveals, preferences* (Ariely and Norton 2008).

One might wonder why in the case of appraisals of one's romantic relationship the most recent attitude is not predictive of one's future behaviour whereas in the case of the preference for faces the most recent choice is reflected in one's future commitments. If reasons cement preferences, then the most recent attitude towards the relationship should be the one shaping behaviour as it was defended with reasons in the course of the experiments.

Here is something that might explain the discrepancy. First, the context of the appraisals seems to matter. In the relationship study, people evaluate one important,

defining aspect of their lives (the relationship with the person they are dating), which affects their wellbeing and their conception of themselves. It is overwhelmingly plausible that participants came to the experiment with pre-existing attitudes about their romantic relationships and romantic partners.

In the faces study, people express a preference about the attractiveness of a stranger's face, which they have never seen before and they will likely never see again. Thus, participants' choices are just an expression of their aesthetic preferences as applied to people's faces, and given the set-up of the experiment, it is very unlikely that there were pre-existing attitudes towards the attractiveness of those faces. It is possible then that the reason-giving exercise cements attitudes and preferences about an object of evaluation when there are no pre-existing attitudes or preferences about that object of evaluation, and the agent has no personal investment in the evaluation itself.

We shall come back to the idea that reasons cement preferences in Sects. 4 and 5.

4 Moral and political preferences

To sum up, we already saw that in the typical choice-blindness scenario, a person is asked to choose between A and B, and chooses A. Moments later, she is asked to give reasons for her choice. Due to a subtle manipulation, her choice is presented to her as having been B. In most cases, the person does *not* detect that the object of her choice has been switched and goes on to provide reasons for choosing B.

Interestingly, the way agents offer reasons for choice B does not differ significantly from the way they offer reasons for their choices when there is no manipulation.[7] In choice error, agents make a mistake as to what their choice is (they say it is B but it really is A). The request for an explanation of why they chose B leads agents to misidentify their choices and give reasons for B. In choice change, agents know what their current choice is, B, but fail to realise the B was not their choice all along. What the experiments show is that agents are neither aware of the instability of their choices nor of the manipulation that caused their choices to change.

The choice change interpretation seems to fit well with choice blindness applying to moral and political statements people are asked to agree or disagree with. In this context, the failure to detect the manipulation causes greater concern than in the context of consumer choice or aesthetic judgements. It is perhaps understandable that participants' preferences about the flavour of jams or the attractiveness of strangers' faces can be easily manipulated, because people may not come to make those choices with existing preferences and may not care very much about the quality of jams or attractiveness comparisons among faces never seen before. However, preferences about moral and political issues just before a major election (e.g., "It is more important for a society to promote the welfare of the citizens than to protect their

[7] The way the agent provides reasons is understood across five dimensions in a comparative linguistic analysis, specifically: *uncertainty, specificity, emotionality, deceit,* and *complexity* (see Johansson et al. 2006: pp. 678–84 for discussion).

personal integrity") are expected to be more stable, and they may even count as self-defining for those agents who see themselves as engaged in politics.

Participants in the studies by Hall and colleagues (Hall et al. 2012, 2013) were asked to fill in what the authors describe as a 'self-transforming' questionnaire on either foundational moral principles (condition one) or on currently topical moral issues (condition two). This happened just before a general election in Sweden, at a time when even people who were not so interested in politics were likely to think about these kinds of issues in their daily lives. Participants had to rate their agreement with a statement using a 9-point scale, and then explain their ratings to the experimenter. The transforming part of the experiment was that two of the statements read out by the experimenter were actually the reverse of the statements originally rated. The rating given was kept the same, but the statement was reversed, so participants were in effect presented with the opposite of the opinion they expressed earlier. The experiment was designed to see whether participants would be led to endorse a view that was in opposition to the one they had just stated.

Once the participant had read the reversed statement, an experimenter would summarise their view back to them with a question: "So you don't agree that [statement]?" or "So you do agree that [statement]?". This mechanism was in place to ensure participants were sure of what they were committing themselves to. The manipulated trials were understood as *corrected* when participants noticed something strange immediately (*spontaneous* detection) or claimed that something was amiss only later, at the time of debriefing (*retrospective* correction). Trials were understood as *accepted* when the participant showed no sign of having noticed that the reversal of the opinion they originally expressed was being fed back to them.

In the first condition, in which the questionnaire was on foundational moral principles, around a third of the trials were spontaneously detected, and a further 8% were retrospectively detected after the experiment. In condition two, nearly 50% of the manipulations were spontaneously detected, but very few participants claimed to detect the manipulations retrospectively. Framed for individuals, 69% of participants *accepted* at least one of the two reversed statements (Hall et al. 2012, p. 4). The manipulation was performed very subtly, so even participants who noticed something strange did not detect the manipulation as such but declared that they must have previously misread or misunderstood the statement.

Interestingly, there was no correlation between self-evaluation of strength of moral conviction and correction, so those "participants who believed themselves to hold strong moral opinions in general were no more likely to correct the manipulations" (Hall et al. 2012, p. 3). And though there was a positive relationship between level of agreement and spontaneous detection, "a full third (31.3%) of all manipulated trials rated at the endpoints of the scale (1 or 9) remained undetected, which shows that not even extreme levels of agreement or disagreement with statements guarantees detection" (Hall et al. 2012, p. 3). However, those participants who claimed to be politically active were more likely to spontaneously detect the manipulation in condition two as compared with politically active participants in condition one. From this we learn that those who identified as politically active were less likely to be manipulated into misidentifying their attitudes or were less likely to have their attitudes changed. It would be interesting to examine whether there are any

other individual differences that can explain the correction/acceptance rates among participants.

In the choice change interpretation, the participant changes her moral or political attitude as a result of the experimental manipulation. The participant really believes that she scored the moral or political statement read to her in the second stage of the experiment. She endorses that attitude and offers genuinely held reasons for it. But what can explain the result that the manipulation is more likely to be detected by the politically active participants? If Lopes is right that second-order beliefs about original choices determine preferences, then in detection cases one of two things might be occurring. The first is that politically active participants are resistant to forming a false second-order belief about their initial preference ("I chose B"), perhaps because they came to the experiment with fixed attitudes about the issues at hand. Alternatively, participants form the second-order belief about the preference they originally stated, but this does not change their preferences, again, because of their having fixed attitudes about the issues at hand. Either way, their behaviour can be explained by the fact that *their political attitude was already cemented at the point of the original choice* and thus resistant to manipulation. People who are politically active may be more likely than people who are not to find themselves in situations where they give reasons for their attitudes prior to participating in the experiment. If it is true that reasons cement preferences, then this might explain why they do not experience choice change.[8]

The conclusion Hall and colleagues draw from the outcomes of the self-transforming questionnaire is that people change in the course of the experiment. Opinions can be "instantly [reversed]" through change blindness.[9] Similarly to the case of attitude shifts we reviewed in Sect. 3.2, the experimental results can be seen as a demonstration of the extreme instability of people's attitudes. The agent's attitude shifts between the time when she is first asked to rate her agreement with the moral or political statement and the time when she is asked to provide reasons for her rating. The change is radical and happens very quickly. What are the implications for self-knowledge? The agent preserves her capacity to identify her attitudes but remains unaware of the fact that her current attitudes were not her attitudes all along. There is no obvious reason to believe that the agent's behaviour indicated a failure of self-knowledge if we identify self-knowledge with correct self-attribution of attitudes. Negative implications for self-knowledge would apply only if we thought that self-knowledge depended on the stability of the agent's preferences and their resistance to manipulation, which would make for a very demanding account of self-knowledge, extending the notion considerably beyond correct self-attribution.

Different from the choice-blindness experiments on consumer choice and faces, the interpretation focusing on the instability of preferences in the case of moral and

[8] It is important to notice that, in later applications of the choice-blindness paradigm to political preferences, the result that politically active participants are more resistant to manipulation has not been replicated (see Strandberg et al. 2018) and political activism made no difference to manipulation detection.

[9] They suggest that "moral decision or judgment is reached through intuition, and the reasons or arguments for the position are mainly constructed through post hoc confabulation" (Hall et al. 2012, p. 5).

political attitudes is at least implicitly endorsed by the experimenters. They imply that in the course of the experiment a new preference is formed due to the manipulation and they also draw some interesting practical conclusions from this for the reliability of opinion polls and the nature of political debate.

5 The advantages of the preference change interpretation

Preference change sounds like a more plausible interpretation of the empirical evidence gathered in some of the choice-blindness studies than choice misattribution. We are not in a position to offer an argument for the general conclusion that choice change should replace choice error in every case, as we suspect that the results obtained under the choice-blindness paradigm resist one overarching interpretation. However, here we offer reasons in favour of choice change that apply to at least some cases of choice blindness.

A first reason in favour of preference change is methodological. It is preferable to take people's reports seriously unless there are good reasons to challenge them. This does not mean, of course, that people are infallible when they talk about their mental lives. Rather, it means that, if they report with apparent sincerity to prefer one option to another and are prepared to defend the option they claim to prefer with reasons, then it is implausible, all else equal, to think that they are massively deceived. If a person tells us that she prefers cinnamon-apple jam to grapefruit jam and gives us good reasons for her preference, on what basis should we challenge her report? The fact that she expressed a different preference earlier is not a good-enough reason to doubt the authenticity of her attribution. After all, we tolerate preference change outside of the laboratory (or the supermarket), even when the person doesn't know why her preferences changed.

A second reason for preference change comes from the evidence reviewed above. The option agents claim to prefer when they are asked to defend their choice with reasons is predictive of their future behaviour. This suggests that the preference change interpretation is not merely supported by the participants' *verbal* reports, but also by the way in which the preference self-attributed in that report, and defended with reasons, shapes their future *verbal and non-verbal* behaviour. The report proves to be more than a momentary glitch in people's self-awareness. This also suggests that manipulating preferences can have lasting effects which is concerning. In the choice-blindness literature people's future behaviour is aligned with their new attitudes or preferences which is evidence that their new attitudes and preferences are not just 'empty talk'. One of the most recent choice-blindness studies (Strandberg et al. 2018) examines in detail the longevity of the preferences induced by the experimenters using the self-transforming questionnaire. Participants are led to form new preferences about health, education, or the environment, and a relationship is found between the persistence of the new preference and the amount of reason giving the agent engaged in at the time of justifying the alleged choice.

How does this happen? Participants were asked to rate a political statement (e.g., "All elementary school students should be offered free homework assistance regardless of their performance and family situation") and then in some cases they

were given the false feedback that their rating (e.g., strongly agree) applied to the opposite statement (e.g., "No elementary school students should be offered free homework assistance regardless of their performance and family situation"). Of the manipulated trials, about 50% were not detected and so the participants did not correct the false feedback and accepted the rating that was reported back to them by the experimenter. Participants were divided into two groups: in one group they just acknowledged the rating (*Acknowledge Condition*), in the other group they were asked to provide reasons for it (*Confabulation Condition*). The attitude change lasted when participants were asked about it straight after the manipulation, and also when they were asked about it a week after the manipulation. In both cases, the attitude change was much more pronounced if the participants had been in the Confabulation Condition. No significant attitude change was observed for those statements for which the feedback had not been manipulated. This confirms the idea that reasons cement preferences.[10]

> On the one hand, absent any manipulation, participants gave the same responses throughout the experiment, clearly indicating they had a stable set of political attitudes. On the other hand, the same participants exhibited large lasting attitude shifts after having accepted the false feedback. (Strandberg et al. 2018, p. 1395)

In addition to the reasons above, we should also think about how the results of the choice-blindness studies, interpreted in the light of choice change, fit into psychological research on the determinants of human behaviour more generally. Although the ease with which agents are manipulated into changing their choices is surprising, it is definitely not surprising that agents' attitudes are subject to change.

Years of psychological research in different research programmes have demonstrated that environmental cues have a more significant effect on our behaviour than we realise. Some obvious examples powerfully illustrating this include the controversial literature on situationism and obedience to authority (Darley and Batson 1973; Milgram 1963). Suggesting that agents may change their choices due to manipulation is compatible with a wealth of independent research pointing to the elusiveness and fluidity of the self. Our selves are *fluid* in the sense that it is illusory to believe that we have fixed personality traits, strong values, or stable attitudes and preferences that are reliably predictive of our behaviour. Further, our selves are *elusive* because their fluidity makes it especially hard for us to have accurate beliefs about what our selves are like.

[10] One might wonder to what extent the choice-change interpretation is compatible with the thesis that reasons cement preferences. Choice change tells us that agents are vulnerable to manipulation in some choice contexts: agents can be made to change their choices without becoming aware of such changes. The thesis that reasons cement preferences does not tell us that agents cannot change their choices but simply tells us that in some choice contexts agents' choices become more stable and more resistant to change than in other contexts. What makes a difference is whether agents provide reasons for the choices they make. Providing reasons for one's choice seems to cement the preferences underlying that choice, effectively *change-proofing* one's choice. More research is needed to ascertain in what circumstances and to what extent reasons cement preferences.

6 Conclusions and implications

We saw that there are two competing accounts of the results of the choice-blindness experiments. What interpretation we favour will determine how we see choice blindness. If choice error is correct, and choice blindness is primarily a mistake in self-ascription, then people can be genuinely blind to their choices and this represents a fairly dramatic departure from self-knowledge. If choice change is correct, and people have dual choices, then people are blind to what causes them to change their minds about things. They do not realise that their preferences can be easily influenced by external factors and they assume that their preferences are more stable than they actually are. Choice error and choice change converge on the fact that agents are misled about the rationality of their choice-making process. However, unless we believe that in order to have self-knowledge people need not only to make correct self-ascriptions, but also to have stable preferences or be resistant to manipulation, self-knowledge is not threatened by choice change.

Why should we take choice change seriously? First, it is a good policy to take people's reports at face value unless we have good reasons to believe that they intend to deceive us. If a person gives us good reasons to endorse her choice of cinnamon-apple jam over grapefruit jam, on what basis should we challenge her report? The fact that she made an inconsistent choice earlier is not a good enough reason to doubt the authenticity of her attribution.

Second, even if we were in the business of discriminating between 'real' and 'fake' choices, it is significant that in the study examining the choice-blindness effect on faces and moral and political preferences people's future behaviour is aligned with their most recent attitudes or choices which is evidence that their most recent attitudes and choices are not just temporary effects of the experimental manipulation.

In addition to the reasons above, we should also think about how the results of the studies on the choice-blindness effect, interpreted in the light of choice change, fit into psychological research on the determinants of human behaviour more generally. Although the ease with which people are manipulated into changing their attitudes and choices is surprising, it is definitely not surprising that people's attitudes and choices are subject to change. Psychological research has demonstrated that environmental cues have a more significant effect on our behaviour than we realise. The influence of such cues would have been difficult to identify in the wild but is obvious in the lab, as the controversial literature on situationism and obedience to authority powerfully illustrates (Darley and Batson 1973; Milgram 1963). This has led to the conclusion that behaviour in general and moral practices in particular are mainly due to features of the agent's situational context as opposed to the agent's character (e.g., Doris 2002). Thus, the idea that people may change their choices due to manipulation is compatible with a wealth of independent research pointing to the fluidity of the self. Our selves are fluid in the sense that it is illusory to believe that we have fixed personality traits, strong values, or stable attitudes and preferences that are reliably predictive of our future behaviour.

If choice change is the correct interpretation of at least some of the psychological studies we reviewed, what are the implications for the centrality of choice in our self-conceptions? Our suggestion is that we should all strive to be better informed about the subtle cues that may influence or even determine choices and attitudes, so as to regain some control over the way in which, and the extent to which, our choices and attitudes can change. If we discover that being *told* that we chose cinnamon-apple jam makes us very likely to *believe* that we chose cinnamon-apple jam and to genuinely endorse that choice as ours in the future, then it is in our interests to pay more attention to, and sometimes challenge, attributions of choices to ourselves made by third parties. After all, those attributions have been shown to have serious power: they can be a device to make us *disregard* and even *reverse pre-existing preferences.* It may not matter when the choice concerns something as trivial as the flavour of jams, but it is much more significant when our moral values and political preferences are at stake.

Psychological research highlighting incoherence and instability in our behaviour can be disheartening. However, it is also empowering, and it is in this spirit that we have argued for the choice change interpretation. Understanding what goes on in the experiments on the choice-blindness effect gives us insight into the measures we can take to enhance our stability and coherence as agents. Understanding the ways in which our preferences are likely to shift as a result of manipulation gives us the means to compensate for our limitations as agents by becoming more aware of those situations where such limitations can be taken advantage of. Moreover, understanding that reason giving contributes to cementing our attitudes and choices gives us further opportunities to ensure that our most cherished commitments will shape our choices in the future. Verbalising and sharing reasons for our commitments render such commitments more stable and make us less vulnerable to manipulation.

Acknowledgements The authors acknowledge the support of the European Research Council under the Consolidator grant agreement number 616358 for a project called *Pragmatic and Epistemic Role of Factually Erroneous Cognitions and Thoughts* (PERFECT). The authors are grateful to two anonymous referees for many helpful comments on an earlier version of the manuscript, and to Petter Johansson and Lars Hall for introducing them to the fascinating phenomenon of choice blindness and for being open to discussing key aspects of their experimental work.

References

Ariely, D., & Norton, M. I. (2008). How actions create—Not just reveal—Preferences. *Trends in Cognitive Sciences, 12,* 13–16.

Bortolotti, L. (2009). The epistemic benefits of reason giving. *Theory & Psychology, 19*(5), 1–22.

Bortolotti, L. (2018). Stranger than fiction: Costs and benefits of everyday confabulation. *Review of Philosophy and Psychology, 9*(2), 227–249.

Carruthers, P. (2005). *Consciousness: Essays from a higher-order perspective: Essays from a higher-order perspective*. Oxford: Clarendon Press.

Darley, J. M., & Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology, 27,* 100–108.
Doris, J. M. (2002). *Lack of character: Personality and moral behavior*. Cambridge: Cambridge University Press.
Hall, L., & Johansson, P. (2008). Using choice blindness to study decision making and introspection. In P. Gardenfors & A. Wallin (Eds.), *Cognition: A smorgasbord of cognitive science* (pp. 267–283). Nya Doxas: Nora.
Hall, L., & Johansson, P. (2009). Choice blindness: You don't know what you want. *New Scientist, 2704,* 26–27.
Hall, L., Johansson, P., & Strandberg, T. (2012). Lifting the veil of morality: Choice blindness and attitude reversals on a self-transforming survey. *PLoS ONE, 7*(9), e45457.
Hall, L., Johansson, P., Tärning, B., Sikström, S., & Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition, 117*(1), 54–61.
Hall, L., Strandberg, T., Pärnamets, P., Lind, A., Tärning, B., & Johansson, P. (2013). How the polls can be both spot on and dead wrong: Using choice blindness to shift political attitudes and voter intentions. *PLoS ONE, 8,* e60554.
Hirsten, W. (2005). *Brain fiction: Self-deception and the riddle of confabulation*. Cambridge: MIT Press.
Ivan, L. (2013). Introspection on romantic relation generates attitudinal change. *Procedia—Social and Behavioral Sciences, 78,* 370–374.
Johansson, P., Hall, L., & Sikström, S. (2008). From change blindness to choice blindness. *Psychologia, 51,* 142–155.
Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science (New York, N.Y.), 310*(5745), 116–119.
Johansson, P., Hall, L., Sikström, S., Tärning, B., & Lind, A. (2006). How something can be said about telling more than we can know. *Consciousness and Cognition, 15,* 673–692.
Lawlor, K. (2003). Elusive reasons: A problem for first-person authority. *Philosophical Psychology, 16*(4), 549–564.
Lopes, D. (2014). Feckless Reason. In G. Currie, M. Kieran, & A. Meskin (Eds.), *Aesthetics and the sciences of mind*. Oxford: Oxford University Press.
Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.
Moran, R. (2001). *Authority and estrangement*. Princeton: Princeton University Press.
Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.
Scaife, R. (2014). A problem for self-knowledge: The implications of taking confabulation seriously. *Acta Analytica, 29*(4), 469–485.
Seligman, C., Fazio, R. H., & Zanna, M. P. (1980). Effects of salience of external rewards on liking and loving. *Journal of Personality and Social Psychology, 38,* 453–460.
Strandberg, T., Sivén, D., Hall, L., Johansson, P., & Pärnamets, P. (2018). False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General, 147*(9), 1382–1399.
Sullivan-Bissett, E. (2015). Implicit bias, confabulation, and epistemic innocence'. *Consciousness and Cognition, 33,* 548–560.
Turner, M., & Coltheart, M. (2010). Confabulation and delusion: A common monitoring framework. *Cognitive Neuropsychiatry, 15*(1), 346–376.
Wilson, T. D., Dunn, D. S., Bybee, J. A., Hyman, D. B., & Rotondo, J. A. (1984). Effects of analysing reasons on attitude-behaviour consistency. *Journal of Personality and Social Psychology, 47,* 4–16.
Wilson, T. D., & Kraft, D. (1993). Why do we love thee? Introspections about dating relationship on attitudes toward the relationship. *Personality and Social Psychology Bulletin, 19*(4), 409–418.
Wilson, T. D., Kraft, D., & Dunn, D. S. (1989). The disruptive effect of explaining attitudes. *Journal of Experimental Social Psychology, 25,* 379–400.