Acta Psychologica 184 (2018) 46–63

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy

Communicated beliefs about action-outcomes: The role of initial confirmation in the adoption and maintenance of unsupported beliefs☆

Toby D. Pilditch[a,*], Ruud Custers[b]

[a] University College London, UK
[b] Utrecht University, Netherlands

ARTICLE INFO

Keywords:
Confirmation bias
Order effects
Communicated beliefs
Instruction effects
Active learning
Probabilistic reversal learning

ABSTRACT

As agents seeking to learn how to successfully navigate their environments, humans can both obtain knowledge through direct experience, and second-hand through communicated beliefs. Questions remain concerning how communicated belief (or instruction) interacts with first-hand evidence integration, and how the former can bias the latter. Previous research has revealed that people are more inclined to seek out confirming evidence when they are motivated to uphold the belief, resulting in confirmation bias. The current research explores whether merely communicated beliefs affect evidence integration over time when it is not of interest to uphold the belief, and all evidence is readily available. In a novel series of on-line experiments, participants chose on each trial which of two options to play for money, being exposed to outcomes of both. Prior to this, they were exposed to favourable communicated beliefs regarding one of two options. Beliefs were either initially supported or undermined by subsequent probabilistic evidence (probabilities reversed halfway through the task, rendering the options equally profitable overall). Results showed that while communicated beliefs predicted initial choices, they only biased subsequent choices when supported by initial evidence in the first phase of the experiment. Findings were replicated across contexts, evidence sequence lengths, and probabilistic distributions. This suggests that merely communicated beliefs can prevail even when not supported by long run evidence, and in the absence of a motivation to uphold them. The implications of the interaction between communicated beliefs and initial evidence for areas including instruction effects, impression formation, and placebo effects are discussed.

Human beings, like the majority of animals, have the capacity to learn how to interact with an environment through first-hand experience of action-outcome relationships. Although some animals have developed the limited ability to communicate these relationships, such as primates, dolphins and bees (Bradbury & Vehrencamp, 1998; Frisch, 1950), humans have taken this ability to much higher levels. This transfer of knowledge can be highly adaptive - we can for instance be informed that having a coffee will cause us to feel more awake, and from this information choose to have a coffee to realize this outcome, without having to start from scratch in working out what might reduce our tiredness. Hence, the development of language has allowed us to transfer information about action-outcomes with an unparalleled capacity and flexibility.

However, despite this communicative capacity, people still seem to hold erroneous beliefs (e.g. the unsupported belief that vaccines cause autism, or homeopathy), whether due to misinterpretations or perceptions of evidence in the communicator, or wilful deception. This combination of erroneous or unsupported beliefs, and the capacity to transfer (a capacity that is ever-increasing with the development of technology, from the printing press to most recently the internet) creates dangerous, viral effects (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012), such as believing an otherwise treatable disease should instead be treated homoeopathically. Such phenomena provoke an obvious and critical question; why are such fallacious beliefs adopted and maintained?

In the present paper, we provide one possible explanation. We argue that when the truth value of a communicated belief is unclear, people use experienced evidence to validate the belief. We demonstrate that in such cases, evidence that is initially encountered will determine whether a belief is consolidated or not, leading to potential bias when this initial evidence is not representative in the long run. Consequently, we believe the present work to be of particular relevance to the literature on persuasion (Briñol & Petty, 2009; Petty & Cacioppo, 1984; Wood, 2000), source credibility (Briñol & Petty, 2009; Hahn, Harris, & Corner, 2009), and instruction effects (Doll, Jacobs,

☆ This work was supported by the Economic and Social Research Council [grant number ES/J500185/1]. Our thanks to Miguel A. Vadillo and Adam J.L. Harris for their help and advice during manuscript preparation.

* Corresponding author at: 26 Bedford Way, London WC1H 0AP, UK.
E-mail addresses: t.pilditch@ucl.ac.uk (T.D. Pilditch), r.custers@uu.nl (R. Custers).

http://dx.doi.org/10.1016/j.actpsy.2017.04.006
Received 30 November 2016; Received in revised form 24 March 2017; Accepted 18 April 2017
Available online 04 May 2017

Sanfey, & Frank, 2009; Liefooghe, De Houwer, & Wenke, 2013; Liefooghe, Wenke, & De Houwer, 2012; Mertens & De Houwer, 2016; Roswarski & Proctor, 2003; Van Dessel, De Houwer, Gast, & Smith, 2015; Van Dessel, Gawronski, Smith, & De Houwer, 2016), given their focus on the impact of *communicated* information.

1. Learning via communication

While information about action-outcome relations has been widely regarded to be represented in terms of associations (Hommel, Müsseler, Aschersleben, & Prinz, 2002). This does not necessarily mean that such representations are always formed by slow associative processes (i.e., Hebbian learning), which are, for instance, thought to underlie habit formation (Custers & Aarts, 2010). They can also result through propositional processes (Mitchell, De Houwer, & Lovibond, 2009), including deduction, inference, and instruction. These allow for fast and flexible changes in associations as these propositions are hypotheses about the state of the world that have a "truth value" and can therefore be confirmed or disconfirmed. Hence, while people may form action-outcome representations slowly through repeated experiences, or via deductive and inferential processes, they may also evaluate the truth value of beliefs about these relations that are communicated by others.

Normative accounts, such as the Bayesian approach, argue that such as evidence is experienced, the belief (and its truth value) is updated to eventually reflect the "true" state of the evidence (Fischhoff & Beyth-Marom, 1983). Within such an approach, a communicated belief, if regarding a new hypothesis, may be considered a "prior". If such a prior is not reflected by the distribution of evidence (i.e. the belief is erroneous), then with sufficient evidence, the effect of the prior would be gradually overruled by experienced evidence. Critically, this highlights the two, interlinked elements that might explain recipients still possessing an erroneous belief: Either the recipient is yet to experience *sufficient* evidence, or the individual is *overconfident* in the prior (although the latter makes the former more likely). Importantly, Bayesian accounts would predict, provided sufficient evidence, that not only should beliefs converge on the "truth" (dictated) by evidence, but that once converged, beliefs should remain there.

However, humans have been found to deviate from this normative standard of learning. Research into cognitive biases has instead shown systematic misinterpretations of evidence (Bar-Eli, Avugos, & Raab, 2006; Gilovich, 1983; Gilovich, Vallone, & Tversky, 1985; Tversky & Kahneman, 1971), and failures to adjust beliefs accurately (Abbott & Sherratt, 2011; Dave & Wolfe, 2003; Dennis & Ahn, 2001; Rozin, Millman, & Nemeroff, 1986; Tversky & Kahneman, 1973) across many domains of learning (for a review, see Pohl, 2004). In particular regard to erroneous belief maintenance, one explanation is an overweighting of belief-congruent evidence, known as a confirmation bias (Klayman, 1995; Nickerson, 1998).

2. Communication and confirmation bias

Confirmation bias is an umbrella term that covers a number of both cognitive and motivational processes (Hahn & Harris, 2014). The impact of these processes is functionally equivalent in terms of the topic of the present paper; it is the retention of an erroneous belief through the overweighting of belief-congruent evidence. We now briefly highlight some of these (at times competing) motivational and cognitive explanations, with a view to demonstrate the importance of assessing the impact of beliefs in the absence of such motivations and cognitive strategies. In doing so, we forward an account of confirmation bias in (erroneous) belief maintenance that is at its heart a consequence of an asymmetry in the way evidence is *integrated*. This integrative bias occurs irrespective of directional motivation (e.g. Kunda, 1990) or skewed evidence exposure (see Klayman & Ha, 1987; Nickerson, 1998) explanations commonly associated with erroneous belief acquisition. Such effects are instead shown to be dependent upon evidence order in the immediate attempted validation of the belief.

2.1. Motivated reasoning

Research in motivated reasoning has argued that *directional* motivations, such as social conformity (Asch, 1955; Cialdini & Goldstein, 2004) and self-concept preservation (Cialdini & Trost, 1998) play a role in confirmation bias effects (Klein & Kunda, 1989; Kunda, 1990). For example, were asked to evaluate the effectiveness of arguments either in favour of, or opposed to, the death penalty (Lord, Ross, & Lepper, 1979). Participants pre-existing political, ethical, and social motivations behind their particular opinion, led to more positive evaluations of arguments that favoured their prior opinion. This was taken as evidence that people are motivated to uphold their personal beliefs when evaluating arguments.

When focusing on the effects of communicated beliefs regarding action-outcome relationships, many of these directional motivations contribute to the confirmation bias effect (Klayman, 1995; Pyszczynski & Greenberg, 1987) in a complex fashion that raises problems for an experimental setting. That is, a communicated belief (e.g., a homeopathic medicine works) may bias evidence integration because it interacts with other needs (such as self-preservation). In other words, the resulting confirmation bias may not directly reflect the communicated belief, but be motivated by the individual's associated needs. Although motivations may attribute to greater degrees of bias, and granted the difficulty in removing all elements of motivated reasoning from real world situations (Yarritu, Matute, & Vadillo, 2013), we posit that merely hearing about a belief is enough to bias evidence integration. Accordingly, such an argument rests on a cognitive explanation.

2.2. Cognitive account

How could a communicated belief lead to confirmation bias effects even in the absence of these motivations? The removal of directional motivations can help clarify the remaining mechanisms at the heart of belief biasing effects. Such a removal has been posited, through work investigating the interaction between motivated reasoning and cognitive processes (Hahn & Harris, 2014; Kunda, 1990), to result in less use of sub-optimal cognitive processes, which might otherwise be selectively employed to favour the motivated outcome. These (biasing) processes can be divided into two camps, first order (or *input* based) and second order (or *integration* based) accounts (MacDougall, 1906).

First order accounts of confirmation bias can be categorized in terms of selective *choices*, such as positive test strategies (Klayman & Ha, 1987; Wason, 1960), selective search (an asymmetry in the scrutiny applied to arguments; see Lord et al., 1979) based, or natural asymmetries in exposure, such as illusory correlations (Fiedler & Freytag, 2004; Fiedler & Krueger, 2011). In all such cases, as an individual learns action-outcomes from experiences, if the evidence *seen* favours confirmation (whether through purposeful strategy, or a naturally skewed environment), any resultant bias could be in part (or entirely) due to this asymmetry in evidence exposure. In other words, if selective information intake is possible within an environment, one cannot discern whether the biasing effect of a communicated belief is due to an asymmetry in the *valuation* of confirmatory evidence over contradictory (Klayman, 1995), or due to the asymmetrical *exposure* to confirmatory evidence (or a combination of the two). Importantly, if selective exposure is the result of one's own actions (rather than pre-determined by the environment), it can be argued that the asymmetry of selection is due to the asymmetry in *evaluation* (i.e. integration; Klayman, 1995; MacDougall, 1906). Accordingly, by precluding selective exposure explanations, it is possible to determine if confirmation bias effects in erroneous belief maintenance may depend upon the skewed *integration* of evidence alone.

Second order (integrative) accounts of confirmation bias have been

seen by theorists as hierarchically responsible for selective strategy use (Klayman, 1995; MacDougall, 1906). The integrative account of confirmation bias posits that confirmatory evidence is *valued* asymmetrically over contradictory evidence. Put another way, despite both confirmatory and contradictory evidence being integrated, the former is systematically over-weighted relative to the latter. Evidence for the over-weighting of confirmatory and underweighting over contradictory evidence has been found in work on confirmation bias (Gilovich, 1983; Klayman, 1984, 1995) and information distortion (Nurek, Kostopoulou, & Hagmayer, 2014). Such work has found support from reinforcement learning (Decker, Lourenco, Doll, & Hartley, 2015; Doll, Hutchison, & Frank, 2011; Doll et al., 2009; Staudinger & Büchel, 2013) and neuroimaging studies (Whitman et al., 2015). The latter of which has demonstrated such an integrative bias is a consequence of the asymmetry between the updating signal that occurs when evidence matches an established pattern of neural connections versus the signal from evidence-pattern mismatches. Put another way, confirmatory evidence confirms (and updates) an established pattern, whilst contradictory (or pattern mismatched) evidence has no equivalent pattern to update (i.e. there is no equivalent "null" pattern that incongruent evidence can update in kind).

When hypotheses are self-generated, there is an initial period of sensitivity as the hypothesis is formed and tested, leading to primacy effects. Evidence for this been found in judgements of causal strength (Dennis & Ahn, 2001; Fugelsang & Thompson, 2003) and information distortion (Blanchard, Carlson, & Meloy, 2014; DeKay, Miller, Schley, & Erford, 2014; Nurek et al., 2014). Alternatively, when an individual is instead presented with a hypothesis second-hand, although the formation process may be absent, there is still an initial sensitivity due to the hypothesis being yet untested. However, unlike self-generated hypotheses, when evaluating a communication, there is an implied relationship to the credibility of its source (Hahn et al., 2009), such that cues indicating the reliability of a source impact the perceived validity of the communication. In this way, source cues may result in circumvention of initial sensitivity.

However, when a belief is communicated in the absence of source cues, then we argue that early experiences, previously demonstrated to be pivotal in hypothesis *formation* processes (Anderson, 1965; Dennis & Ahn, 2001), are instead required to validate the belief. Further, by removing source cues such as affiliation with the source (Frost et al., 2015), and perceived expertise (Goodwin, 2011; Harris, Hahn, Madsen, & Hsu, 2015; Walton, 1997), the resulting motivational and cognitive explanations for confirmation bias (e.g., the belief was communicated by a friend, whom one is motivated to agree with) are reduced. Such a reduction distances the present work from literatures including argumentation (Hahn et al., 2009; Harris et al., 2015), attitude change and persuasion (Briñol & Petty, 2009; Petty & Cacioppo, 1984; Priester & Petty, 1995), in which source cues are themselves taken as evidence with a truth value. Such work has typically indicated the efficacy of an argument (or belief) as dependent upon other truth-value assessments of source cues, which interact with an individual's priors (e.g., a prior attitude or opinion).

3. Truth values, integrative bias, and instruction effects

Propositions, which communicated beliefs may be considered to be, imply a truth value (Strack & Deutsch, 2004). Whilst associations, in their unqualified activation links *are* the state of the world (Shanks, 2007), propositions are statements regarding the state of the world, and can thus differ in their accuracy. How truth values may then interact with evidence is therefore of interest - how do people treat such information, especially if its value is uncertain?

Work in instruction effects has typically looked at the impact of formal instruction (notably from an experimenter source) on automatic processes, such as approach-avoidance (Van Dessel et al., 2016) and task-rule congruency effects (Liefooghe et al., 2012). In the present work, we play off the inherent uncertainty of truth values of communicated beliefs (i.e. instructions) by pitting evidence against the instruction. This introduces new questions that cannot be assessed when evidence and instructions are in line. For example, this allows for the investigation of the impact of *order effects* on instruction efficacy on longer, more deliberative learning processes. The present work seeks to provide a novel contribution in this manner, by demonstrating that the presentation of a hypothesis (or instruction), even *incidentally communicated* (i.e. without complementary directional motivations), shapes how evidence is then integrated over time. Further, we argue that instruction effects do not depend upon directional motivations (such as authority effects).

4. Present research

The approach of the present work bears a parallel to previous research in the Judgement and Decision Making literature, which has focused on the roles of communicated (termed 'description') and experienced evidence as advice-taking (Bonaccio & Dalal, 2006; Harvey & Fischer, 1997). Typically, these paradigms have used binary choices between risky (probability of high reward or nothing) and safe (guaranteed low reward) gambles, in which participants must use either their own experience or on descriptions of the choices provided by the experimenter (Ludvig & Spetch, 2011; Newell & Rakow, 2007; Rakow & Newell, 2010). Although, this field has recently started investigating direct competition between these two forms of information (Weiss-Cohen, Konstantinidis, Speekenbrink, & Harvey, 2016), methods typically incur potential experimenter demand effects (Hertwig & Ortmann, 2008; but for a notable exception, see Yaniv, 2004), which introduces unaccounted for directional motivations. Further, the use of 1:1 description to evidence ratios (each choice re-introduces experimenter instruction, leading to potential over-weighting) and quantified statements (e.g., "60% chance to win $2") rather than more ecologically valid, relational, action-outcome beliefs, (i.e. generic, unquantified statements; Cimpian, Brandone, & Gelman, 2010; Gilovich, 1993; Leslie, 2008) separates this literature from the present research.

Instead, the line of research developed here looks to implant a communicated belief in a manner that allows for uptake in the reduced presence of additional motivations (such as experimenter or authority effects). Through Amazon's Mechanical Turk (MTurk), an on-line lottery context was used in which participants were told they would be choosing between two lotteries repeatedly, trying to generate the greatest overall payment. A novel manipulation was designed that communicated beliefs through an on-line "comment section" prior to the task. Participants were shown anonymous comments from "previous players" regarding the task, under the guise that they could add their own comments once they had finished making their choices, with the aim of providing information to both the task developers and other players. The "previous players" comments were in fact generated by the task itself, with most being neutral in nature, whilst those in the belief conditions contained a directional hypothesis. Participants were taken through filter questions at the end of the experiment to ensure they had not seen through the manipulation. Such a manipulation provided a novel, ecologically current form of communicating a belief that avoids aforementioned motivation pitfalls (Kunda, 1990), as participants believed they were simply playing a game with a monetary incentive.

Following this manipulation was a series of binary choices between the two lottery machines. One of the two machines (unknown to the participants) would start off as the probabilistically dominant option for a number of trials, before these probabilities then reversed, known as a probabilistic reversal (Peterson & DuCharme, 1967). Having a reversal of evidence resulted in three between subject groups: a control group (received no communicated belief), a belief group that received initial supportive evidence (BiS group), and a belief group that received initially undermining evidence (BiU group). All groups saw the same

two-sided evidence. That is, all participants saw exactly the same evidence for the two options (which had equal outcomes overall), only the order of these outcomes was pseudo-randomised to create two "phases" in which one option dominates the other.

4.1. Hypotheses

The design outlined above allows for several predictions to be tested, discerning between potential accounts. Importantly, both the belief manipulation and evidence presentation aim to improve upon previous research by reducing aforementioned directional motivations and selective evidence exposure, both of which add unwanted alternative explanations for any subsequent bias. The use of a reversal additionally allows for the demonstration of ongoing learning. If participants are sensitive to reversals, between-group deviations can be said to result from a biased active learning process, rather than due to a gradual lapse in attention over time to new evidence.

While normative accounts of learning predict that people's choices should be dictated by communicated beliefs at first, they also assume that these beliefs are updated based on experience evidence, so that over time the communicated belief is washed out and the new belief reflects the evidence (i.e. a normative account). There are several ways, though, it which lasting biased could manifest itself. First, if people put too much trust in the communicated belief, even a lot of contradicting evidence may not be enough to erase this strong prior belief (a "strong prior" account). Second, people may put not enough trust in the evidence, failing to update their beliefs appropriately (a conservatism account; see Phillips & Edwards, 1966; Pitz, Downing, & Reinhold, 1967). We argue, however, based on insights from source credibility work (Chaiken & Maheswaran, 1994; Hahn, Oaksford, & Harris, 2012; Harris et al., 2015), order effects in hypothesis testing (Anderson, 1965; Dennis & Ahn, 2001; Nurek et al., 2014), and integrated confirmation bias accounts (Klayman, 1995; MacDougall, 1906; Whitman et al., 2015), that as a communicated belief is likely to require validation, its influence will be dependent upon initially supporting evidence. In this case, effects of communicated beliefs and initial evidence are thought to interact, wherein too much trust is put in the belief once it is confirmed by evidence, leading to a confirmatory bias in evidence integration.

5. Experiment 1

5.1. Method

Following the outline set out in the present research. Experiment 1 was designed using an 80-20 probabilistic reversal with 100 trials each side of the reversal, resulting in 200 trials in total. These trials were preceded by the aforementioned "comment section" which contained a communicated belief for the manipulation groups ("Machine A seemed luckier to me"), with the remainder (and for the control group, the entirety) of the comments of a neutral nature (e.g., "fun task", "seemed interesting").

5.1.1. Participants

Participants were recruited and participated online through MTurk. Those eligible for participation had a 95% and above approval rating from over 500 prior HITs. Participants completed the experiment under the assumption the purpose of investigation was general gambling behaviors when using multiple lotteries. Participants were English speakers between ages 18 and 65, located in the United States. Informed consent was obtained from all participants in all experiments.

5.1.2. Design

Two lottery machines, labelled A and B, that both generated six number outcomes were used as choice context (see Appendix A for an example trial output). The instructions given to participants explained that each machine uses a unique algorithm, based on the hypergeometric distribution of outcomes intrinsic to most modern lotteries (Stern & Cover, 1989).

The order of outcomes between the two machines were structured into two phases of 100 trials. For the first phase 1 machine yielded 80% of the "wins" - classified as providing an outcome better than the alternative (e.g., machine A has one ball that matches the ticket, whilst machine B has two balls that match the ticket, so machine B has "won"), whilst the other yielded "wins" 20% of the time. During the second phase these probabilities reversed. Hence, the overall proportion of outcomes matched the natural probabilities of the hypergeometric distribution found in 6 ball lotteries. Within each phase the order of outcomes was randomized. Which machine started off dominant was counterbalanced between participants. In line with the natural odds, 20% of trials were consequently uninformative as they involved draws (0-0, 1-1, 2-2) between the two machines, the remaining diagnostic trials (80%) followed the aforementioned probabilistic reversal.

5.1.3. Procedure

Before starting the trials, participants were shown a "comment section" in which previous participants had written their thoughts regarding the task. These comments were rigged to appear to be from other MTurk participants (complete with fake MTurk ID numbers), and were of a hypothesis neutral nature ("interesting task.", "good fun, thanks!"). For the belief conditions, instead of all comments being neutral (as in the control condition), the top comment was a communicated belief regarding the two machines ("I felt that machine A(B) was much luckier!").

On each trial participants pressed a button that generated a "ticket" of three numbers, and then chose a machine to gamble with on each trial. Participants were invited to earn as many points as possible, based on the number of matches between their "ticket" and their chosen machine. Each trial cost participants one point, so a failure to match any numbers resulted in a net loss of -1 point, whilst a single ball matched earned 2 points, 2 balls matches earned 8 points, and 3 balls matched earned 50 points, reflected the increasing rarity of these outcomes (see Appendix A for an example feedback screen). Participants were aware of their current total points earned during the trials, and instructed that their total amount of points directly corresponded to an increasing bonus payment in dollars (e.g., passing the 50 point threshold increased the standard payment by 10%, then passing the 100 point threshold increased the payment by a *further* 15% of the standard payment, for a total of a 25% bonus, and so on).

For each trial, once participants had generated their ticket of numbers and selected a machine with which to gamble, participants then pressed a button to generate the outcomes for that trial. Each machine drew 6 balls, numbered from 1 to 49, with a new draw for each trial. Matches between the numbers of the participant and the selected machine were highlighted in green, whilst the forgone matches of the non-selected machine were highlighted in red.

Once participants had completed all gambles, demographics were filled out, along with questions regarding how often the participants felt they had chosen optimally, and how often they felt the other machine had provided a better outcome. Participants were also asked about the probabilities of each outcome for each of the machines, along with a brief questionnaire assessing various known correlates of superstition and gambling behaviours; locus of control (Levenson, 1973), the revised Paranormal Belief Scale (Tobacyk & Milford, 1983) items concerning luck, and neuroticism. Finally, participants completed a series of exit funnel questions to assess awareness of the comment section manipulation. Following completion of the task, participants were debriefed and given an email to contact if they had any further questions.

The main dependent variables under investigation were the proportion of choices made in favour of the initially dominant machine. We hypothesised that the BiS group (who receive initial support for the

belief) would select the initially dominant machine in phase 1 (pre-reversal) significantly more than controls, and further, that this difference would persist into phase 2 (post-reversal). For the BiU group (who do not receive initial support for the belief), several aspects of communicated biasing effects became possible to test:

Firstly, phase 1 could demonstrate whether the communicated belief acts as a strong prior (i.e. the belief is given a very high value that takes a large amount of evidence to over-rule), which would result in the BiU group significantly differing in their proportion of choices relative to controls, favouring their belief indicated machine, in spite of its initial sub-optimality.

Secondly, for phase 2, two possible effects could occur in the BiU group: either the BiU group would favour the now dominant machine (which matches their communicated belief) more so than controls (which leads to a linear effect of condition in phase 2), or the BiU group would have refuted the belief and would be no different from controls in phase 2.

5.2. Results

5.2.1. Descriptives and processing

The 400 participants recruited were US based, randomized into either the BiS (161), BiU (137) or control (102) conditions. The mean age was 35.52 years (48% female). The data were gathered in two stages: an initial run of 80 participants, with just the BiS and control groups allowed for an estimate of the power needed when also adding the BiU group (to test additional predictions) in run 2. This power analysis, using G*power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007) was run using the smallest effect size of the dependent variables of interest, and indicated that to detect a significant effect of condition at the .05 level with 80% power would require an average group size of 91. For the two manipulation conditions this number was multiplied by about 1.5 to compensate for failures in passing the manipulation check (mentioned below), resulting in a total N of 360. Given the unknown nature of the additional BiU group, this was conservatively increased by 10% to 400. An ANOVA analysis was run, using experiment number as a covariate, finding no significant differences of experiment number on all dependent variables.

After completing the task, all participants were asked a series of filter questions to determine whether the cover story (comments originating from other participants, rather than the experimenter) was believed, culminating in a check of whether the comment manipulation had been remembered.[1] If participants had no recollection of a manipulation comment (if one was presented in their condition), they were removed from subsequent analysis,[2] leaving 110 in the BiS group (75% pass rate) and 81 in the BiU group (59% pass rate). The decision to remove those who failed to remember was taken to reduce noise in further analysis break downs, specifically when breaking down the analysis into phases, participants who failed this check added a large amount of variance when analysing at a finer level. Furthermore, by removing those who fail the manipulation check, it is possible to better ensure possible differences between groups in remaining participants (notably for the BiU group) were not due to failures in memory or registering the manipulation in the first place.

The difference in the proportion of those who failed to remember the manipulation between groups was not significant. The remaining 293 participants (49% female), average age 35.12 years (SD = 12.08), were used for the analyses below.

5.2.2. Correlates

Locus of Control, age, gender, gambles per week and revised Paranormal Belief Scale (rPBS) variables were not correlated with any of the dependent variables.[3]

5.2.3. Choice data

The key dependent variables used in the analysis were the total proportion of choices made in favour of the initially dominant machine, and the proportion of these choices made broken down by phase (shown in grey in Fig. 1).

A series of ANOVA tests were conducted to determine the main effects of condition and phase on the proportion of choices made. Significant effects of condition, $F(2, 289) = 5.041$, $p = 0.007$, and phase, $F(1, 289) = 137.84$, $p < 0.001$, were found on the proportion of choices made in favour of the initially dominant machine. The interaction term between phase and condition was not significant.

A series of pairwise comparisons between groups both overall and within each phase was conducted to break down the main effect of condition. These comparisons found the BiS group to be significantly higher than both controls, $F(1, 211) = 4.804$, $p = 0.029$, $\eta^2 = 0.022$, and BiU groups, $F(1, 190) = 8.579$, $p = 0.004$, overall. Furthermore, when broken down by phase, the BiS group was significantly higher than the BiU group in both phase 1, $F(1, 190) = 5.152$, $p = 0.024$, and phase 2, $F(1, 190) = 4.401$, $p = 0.037$, whilst the difference between BiS and controls did not reach significance for phase 1, $F(1, 211) = 1.697$, $p = 0.194$, $\eta^2 = 0.008$, or phase 2, $F(1, 211) = 3.733$, $p = 0.055$, $\eta^2 = 0.018$.

The differences between the BiU and control groups were not significant overall ($p = 0.329$), in phase 1 ($p = 0.318$) or phase 2 ($p = 0.632$),[4] a trend that is evidenced in the phase lines (grey) of Fig. 1, suggesting that the undermined belief is abandoned. Visual inspection of the choice data suggested that the BiU group are no different from controls, whilst the BiS group is significantly different from the two, in line with the hypothesis that biasing effects of a second-hand belief relies on exposure to initially supporting evidence. This was further corroborated by the pairwise comparisons, leading to the development of a post-hoc contrast code analysis.

5.2.4. Post-hoc contrast code formation

Accordingly, to test whether the BiU group were no different than controls in comparison to the BiS group, a contrast code ANOVA was conducted (BiS, Control, BiU: 2, −1, −1). The contrast code analysis of overall choice proportion was significant, $F(1, 289) = 9.535$, $p = 0.002$, $\eta^2 = 0.032$, demonstrating a significantly higher number of choices in the BiS group as compared to both control and BiU groups.

Breaking this down by phase, (as illustrated by the grey lines in Fig. 1), the contrast code persisted in both phase 1, $F(1, 289) = 4.509$, $p = 0.035$, $\eta^2 = 0.015$, and phase 2, $F(1, 289) = 6.077$, $p = 0.014$, $\eta^2 = 0.021$, again demonstrating a significantly higher number of choices in the BiS group as compared to both control and BiU groups. The interaction between phase and contrast code was not significant ($p = 0.676$).

[1] The exact phrasing of this question was “Do you think comments were biased towards one machine?”, with options of A, B, or No. If participants selected “No”, this was deemed grounds for removal. The question was preceded sequential exposure to first the question “Did you notice anything funny about the start of the experiment?” (open text response), to assess if participants suspected the cover story, followed by “Was there anything that influenced you regarding the previous participant comments?” (open text response).

[2] Significant differences were found with all participants included between conditions on the overall proportion of choices, $F(2, 399) = 4.291$, $p = 0.014$, $\eta^2 = 0.021$, and the subsequent contrast coding of the same analysis, $F(1, 396) = 8.185$, $p = 0.004$, $\eta^2 = 0.02$. This pattern of results is explained further in the main effects section.

[3] Due to a minor programming error, one counterbalance condition had a slight imbalance in the number of 2 ball matches in the second phase on one machine. To remove possible issues, counterbalancing was used as a covariate in all further analyses, as the counterbalancing factor was not exactly even across groups.

[4] Bayesian T-tests of these pairwise comparisons were conducted using the JASP statistical programme (JASP Team, 2016), using a uniform prior across possible models (as used across all subsequent Bayesian analyses unless specified otherwise). Substantial support was found for the null for overall, $BF_{10} = 0.249$, phase 1, $BF_{10} = 0.259$, and phase 2, $BF_{10} = 0.178$, in accordance with the $< 1/3rd$ cut off recommendation (Dienes, 2014).

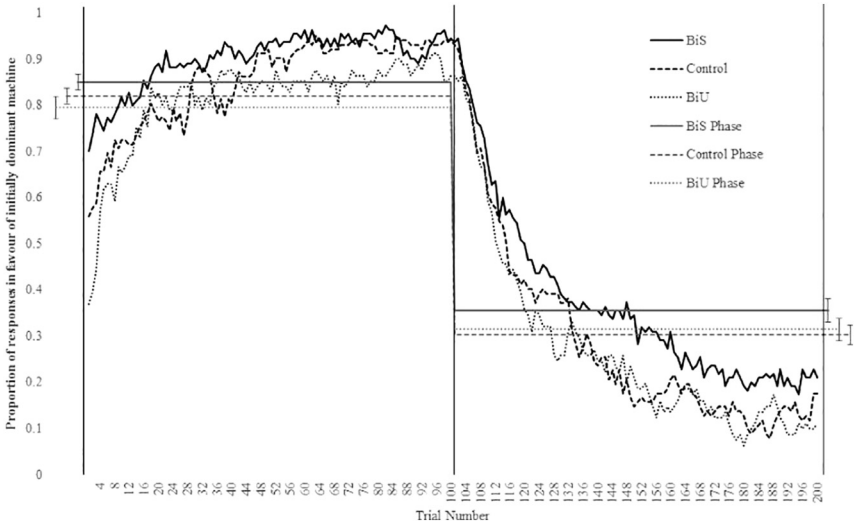


Fig. 1. Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially dominant machine as participants move through the 200 trials (with reversal occurring at 100 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

5.2.5. Phase 1 convergence

Visual inspection suggested that by the end of phase 1 participants had all converged towards the dominant machine. To test this, a repeated measures ANOVA of contrast coded condition (between subjects) × 10 trial epochs (10 in total) within-subjects was conducted. The consequent interaction between epoch and contrast code was significant, $F(2, 290) = 5.657$, $p = 0.004$, $\eta^2 = 0.038$, indicating a convergence of the contrast effect over the course of the phase.[5]

5.3. Discussion

Several conclusions follow from these results. As can be seen from Fig. 1, there is a strong effect of reversal, indicating that learning is ongoing in all groups. As such differences between groups are unlikely to be due to task disengagement and instead suggest participants were attentive to changes in the evidence. It should also be noted that this sensitivity to reversal was present in all three groups, which indicates the phase 2 contrast effect is not due to the BiS group no longer being attentive to any changes in evidence. In line with this, the convergence of all groups during the first phase suggests that the contrast effect during this phase could be due to starting point differences between the BiS and other groups. However, the subsequent re-emergence of the contrast effect in the second phase indicates that the effect of the belief seen in phase one has not been washed out by the evidence, and instead the BiS group's belief is still playing a role in biasing the integration of subsequent evidence, relative to the other groups.

This convergence is one indicator that a conservatism explanation of bias does not fit across the three groups as well as suggesting that a strong prior explanation for the effect of the communicated belief is not appropriate. This leads to the final, and most important conclusion of Experiment 1: even with the presence of 2-sided evidence, and the reduction of motivational explanations of confirmation such as experimenter demand effects (along with the introduction of a competing (against belief adherence) accuracy incentive), it is initial evidence that dictates a belief's subsequent biasing effects. The immediate mapping of the initially undermined (BiU) manipulation group onto the control group suggests that initial evidence is needed to consolidate a communicated belief. It is important to further note that given the manipulation check criteria, the similarity between the BiU group and control group in the analyses was not due to participants in the BiU group having forgotten the communicated belief. This suggests that those in the BiU group have *refuted* their communicated belief, and are thus unaffected by it, even when the evidence changes to support it. Consequently, the most suitable explanation for the effect of a communicated belief on evidence is initially supportive evidence is required to consolidate the belief, but once consolidation has occurred, learning is still active, but biased to favour confirmation when interacting with subsequent evidence.

There are several limitations pertaining to Experiment 1. Firstly, the contrast analysis for Experiment 1 was post-hoc, and therefore requires replication. Secondly, the probability distributions used in the tasks may have been too strong. By having an 80/20 probability distribution, the dominance of one machine over the other was clear to all groups (as evidenced in Fig. 1) demonstrated by their convergence during phase 1. All groups were also able to detect the reversal that occurred at trial 100, which is not unsurprising given the severity of the reversal, as the initially dominant machine becomes 60% worse in combination with the counterfactual showing the dominance of the alternative. This may have led to ceiling effects in the number of choices between manipulation groups and controls, that might have been teased apart better by a more uncertain environment.

Returning to the literature on confirmation bias effects, this may have led to an increased plausibility of alternative hypotheses that allowed manipulation groups to negate the validity of the second-hand belief (Klayman, 1995), muting the efficacy of a bias that might otherwise have persisted under uncertainty.

6. Experiment 2

Accordingly, the purpose of Experiment 2 was to replicate the exploratory contrast effect of Experiment 1 *a priori*, improving upon the design given prior limitations.

6.1. Method

The design and procedure followed that of Experiment 1, with the following changes outlined below.

Firstly, despite the intriguing potential real world ramifications of the effect that such a small intervention can have on updating, a stronger manipulation (increasing the number of comments indicating

[5] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast effect in the first epoch, and the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first epoch, $BF_{10} = 149.4$, whilst strong support was found for the null in accordance with the < 1/3*rd* cut off recommendation for the final epoch, $BF_{10} = 0.136$.

the same directional belief) was constructed. The comment manipulation was increased in strength from one second-hand belief to three (all in the same direction). This was done to improve the rate of manipulation check failures experienced in Experiment 1 by improving the visibility of the manipulation to participants. Additionally, it increased the reliability of the communicated belief (Siegrist, Cvetkovich, & Roth, 2000; Yaniv & Kleinberger, 2000), as there were now several (supposedly independent) sources all providing the same preference. As such, the belief could be interpreted as a trend (hence more likely to be valid), rather than a one-off occurrence. Relative to neutral comments, these manipulation comments were still in the minority, to avoid social conformity issues.

Secondly, the probabilistic reversal was altered from 80/20 - 20/80 to 70/30 - 30/70. This reduction in the severity of the reversal was introduced to help reduce possible ceiling effects discussed above due to the dominant machine (and reversal) being too obvious to both controls and manipulation groups alike. This change was aimed at teasing apart possible biasing effects further.

The third change was the introduction of three posterior measures at the end of the main task, in which participants chose which of the two machines they preferred ("Which machine do you think is better?"; binary preference), how confident they were (0–100%) in that preference, followed by their estimation of the distribution of better outcomes between the two machines ("What is the spread of better outcomes between the two machines?", from 100% A, through 50/50, to 100% B on a 100 point scale). The addition of posterior measures allows for the testing of whether the bias seen in the BiS group in Experiment 1 that converged with the other groups during phase 1, but then re-emerged following the reversal is a reflection of a truly consolidated belief. The measures were therefore included as a supplemental, exploratory measure to investigate if the contrast effects in the main (behavioural) dependent variables are found in end-of-sequence judgements as well (Hogarth & Einhorn, 1992).

6.1.1. Hypothesis

The hypotheses for Experiment 2 are primarily based on the contrast code effects found in Experiment 1. Those who receive a belief that is initially supported (BiS) will choose the machine indicated by the belief significantly more than controls, whilst those in the group that receive initially undermining evidence (BiU) will not show such a bias and be no different from the control group. Furthermore, with the inclusion of the posterior probability measures, contrast effects were predicted to extend to these measures as well.

6.2. Results

6.2.1. Descriptives and processing

Based on the contrast analysis of Experiment 1, a second power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 2. Converting partial eta squared values for the contrast code analyses of the three dependent variables into Cohen's d (Cohen, 1992) effect sizes, using the smallest of these effect sizes, to detect a significant effect at the .05 level, with 80% power resulted in an average estimated sample size of 90 per group. Following the same procedure as Experiment 1, the groups sample sizes were increased by 33% to compensate for those failing manipulation checks, calculated from the failure rates of Experiment 1, resulting in a total N of 330. Given the changes to the paradigm, this was conservatively increased by 10% to 360.

Participants were recruited online using MTurk. Those who had taken part in the previous experiment were ruled out from participating. The 360 participants recruited were US based, randomized into either the BiS (121), BiU (122) or control (117) conditions. The average age was 34.7 years (SD = 11.47) and the sample was 49% female. After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comment (if one had been presented in their condition), they were removed from subsequent analysis,[6] leaving 103 in the BiS group (85% pass rate, up from 75% in Experiment 1) and 96 in the BiU group (79% pass rate, up from 59% in experiment 1). The decision to remove those who failed was taken following the same protocol and reasoning as Experiment 1.

Having increased the number of comments from one to several, and reduced the severity of the probability distributions, the drop-out rate was lower in Experiment 2, although these differences (both between groups and between studies) were not significant. The following analyses were conducted using the remaining 316 participants, with an average age of 34.71 years (SD = 11.41) and 50% female.

6.2.2. Correlates

Locus of Control, age, gender, gambles per week and revised Paranormal Belief Scale (rPBS) variables were not correlated with any of the dependent variables.

6.2.3. Choice data

As can be seen in Fig. 2, the reversal point was harder to detect for all participants, but nevertheless by the point of reversal all groups had learnt a preference for the dominant machine, and subsequently moved in the correct direction upon reversal. The proportion of choices in favour of the initially dominant machine, both overall, and broken down into pre- (phase 1) and post-reversal (phase 2) proportions were the key variables of interest once again for running the contrast code analysis.

To assess whether Experiment 2 replicated the key findings of Experiment 1, the same contrast code analysis procedure was conducted for the three conditions (BiS, Control, BiU: 2, −1, −1), along with pairwise comparisons between groups. The contrast code analysis of condition on the overall proportion of choices made was significant, $F(1, 313) = 13.637$, $p < 0.001$, $\eta^2 = 0.042$. Along with corroborating pairwise ANOVA analyses which show the BiS group was significantly higher than controls, $F(1, 219) = 15.192$, $p < 0.001$, $\eta^2 = 0.065$, and BiU, $F(1, 198) = 7.111$, $p = 0.008$, $\eta^2 = 0.035$, groups, whilst the difference between controls and BiU was not significant ($p = 0.318$).

Breaking this down by phase, the proportion of choices in the first phase (grey lines in Fig. 2, prior to reversal point), the contrast code was significant, $F(1, 313) = 16.69$, $p < 0.001$, $\eta^2 = 0.051$. This was corroborated by pairwise ANOVA showing the BiS group was significantly higher than controls, $F(1, 219) = 19.993$, $p < 0.001$, $\eta^2 = 0.084$, and BiU, $F(1, 198) = 8.188$, $p = 0.005$, $\eta^2 = 0.04$, groups, whilst the difference between controls and BiU was not significant ($p = 0.129$).

These effects continued into the second phase (grey lines in Fig. 2, post reversal point), as the contrast code was again significant, $F(1, 313) = 5.046$, $p = 0.025$, $\eta^2 = 0.016$. This was corroborated by pairwise ANOVA showing the BiS group was significantly higher than the control group, $F(1, 219) = 4.739$, $p = 0.031$, $\eta^2 = 0.021$, whilst the difference between controls and BiU was not significant ($p = 0.805$). This difference between the BiU and BiS groups was not significant in phase 2 ($p = 0.077$), however as the two key comparisons (BiS is different from controls, whilst BiU is not) remain significant, the position of the BiU group proportion (on the BiS group side of the control group) further supports the notion of the BiU group refuting their belief.

[6] Following the same protocol as Experiment 1, significant effects were found for both the effect of condition on the overall proportion of choices, $F(2, 359) = 4.76$, $p = 0.009$, $\eta^2 = 0.026$, and the contrast coding of the same analysis, $F(1, 357) = 8.233$, $p = 0.004$, $\eta^2 = 0.023$, regardless of manipulation check removal.

Fig. 2. Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially dominant machine as participants move through the 200 trials (with reversal occurring at 100 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

6.2.4. Phase 1 convergence

Following the analysis protocol of Experiment 1, a repeated measures ANOVA was conducted to assess the degree of convergence in choice proportions over the course of phase 1. The contrast code was the between-subjects grouping factor, and 10 trial epochs were within-subjects. The consequent interaction between epoch and grouping was significant, $F(2, 313) = 8.941$, $p < 0.001$, $\eta^2 = 0.054$, indicating a convergence of the contrast effect over the course of the phase.[7]

6.2.5. Posteriors

The posterior measure of estimated probability distribution (see Fig. 3) showed a significant contrast code effect of condition, as found in the above behavioural measures, $F(1, 313) = 4.192$, $p = 0.041$, $\eta^2 = 0.013$.[8]

A further series of t-tests revealed a primacy effect, wherein the initial evidence favouring one machine dominated the reversed evidence in the latter half of the task, in the BiS, $t(102) = 5.862$, $p < 0.001$, 95% CI [5.43, 10.99], control, $t(116) = 3.586$, $p < 0.001$, 95% CI [2.3, 7.99], and BiU, $t(95) = 2.129$, $p = 0.036$, 95% CI [0.25, 7.05], groups.

Participant's binary preference posterior measure did not yield a main effect of condition, but did show strong primacy effects, wherein the initial evidence favouring one machine dominated the reversed evidence in the latter half of the task, in BiS, $X2(1, N = 103) = 14.767$, $p < 0.001$, control, $X2(1, N = 117) = 9.308$, $p = 0.002$, and BiU, $X2(1, N = 96) = 12.042$, $p < 0.001$, groups. The confidence measure for the binary preference did not show any significant effects of condition, or order effects.

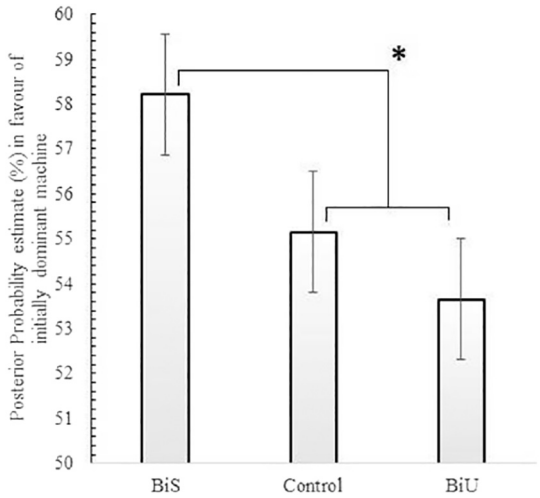


Fig. 3. Posterior probability estimate as a percentage of better outcomes. Greater than 50% reflects a preference for the initially dominant machine, less than 50% indicates a preference for the initially sub-optimal machine. Outcomes split by group (*$p < 0.05$).

6.3. Discussion

Replication of the findings of Experiment 1 demonstrates the necessity of initial supporting evidence in validating the second-hand belief, leading to subsequent biases in updating. Similar to the effects discussed in Experiment 1, phase 1 differences can be seen as due to adjustment from initial starting points (as dictated by communicated belief). However, despite all participants (regardless of group) converging beyond a probability matching (Edwards, 1961) level (choosing the dominant option 70% of the time) by the end of phase 1, the contrast code once again re-emerges in phase 2.

Importantly, when investigating the posterior measures, despite a general tendency towards primacy amongst all groups, which is not surprising for an end-of-sequence judgement (Hogarth & Einhorn, 1992), an asymmetry between the primacy in the BiS group as compared to the other groups existed. This was corroborated by a replication of the contrast analysis for the posterior probability estimate. Such an extension lends credence to the argument that the biasing effect due to consolidating a communicated belief has persistent effects beyond trial-by-trial choices and into end-of-sequence judgements. Furthermore, such a bias occurred despite all groups witnessing

[7] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast code in the first block in direct comparison to the assessment of the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first block, $BF_{10} = 4.73 * 10^5$. In the final epoch, there was no longer support for the contrast effect, but this did not reach the $< 1/3rd$ to be classified as substantial support for the null, $BF_{10} = 1.731$. However, it should be noted that a Bayesian ANOVA on the subsequent epoch (first epoch post-reversal) does indicate a reversal in trend from convergence to divergence upon reversal, as strong evidence is once again found for the contrast effect, $BF_{10} = 10.41$.

[8] A Bayesian T-test was conducted to confirm the lack of a difference between controls and BiU was not due to insufficient power by looking for strong support of the null (a Bayes Factor of $< 1/3rd$). Substantial evidence was found in support of the null, $BF_{10} = 0.186$.

the same evidence (due to the presence of counterfactual feedback), which suggests this difference is not a product of a differential in evidence exposure.

Finally, Experiment 2 replicated the effects found in Experiment 1 using a 70/30 probability reversal, which reduced ceiling effects in the proportion of choices made within each phase. The second change to paradigm - including multiple comments indicating the communicated belief - resulted in a lower proportion of participants failing the manipulation check and likely attributed to a stronger contrast effect, in line with models of source trust in advice-taking literature (Siegrist et al., 2000; Yaniv, 2004), as perceived trust is increased when several sources indicate the same information, up from a single source.

7. Experiment 3

Experiment 3 set out to extend the effects found in Experiments 1 and 2 into the domain of health decision making. The first reason for this change was to ratify possible claims of generalizability of the effects in question, by moving outside of a gambling context. Secondly, in making the transition away from gambling, it was hoped that improvements could be made in short-term "streak-shooting" noise in the choice data. In other words, given the aim of the experiments were to have participants focus on the long-term, overall quality of the two options available to them, the context of lotteries is conducive to short-term, fallacious strategies such as gambler's fallacy (Ayton & Fischer, 2004; Barron & Leider, 2010; Jessup & O'Doherty, 2011). Such strategies are based on recent performance, such as assuming that because one machine has not won recently, then it is "due" for a win. These strategies are hence, in relation to the overall effects under investigation, a possible distraction.

Furthermore, by moving into the domain of health, outcomes are defined as a medicine either "curing" or "failing to cure" the disease in a particular patient. In Experiments 1 and 2, where outcomes were variable (number of ball matches) for each machine, and thus comparisons between machines had to be based on a calculation of the relative number of matches between options (e.g., an assessment of whether a two-ball match is better than 3 sets of one-ball matches).

By simplifying this to a 1 (cure) or 0 (fail to cure) outcome for each option, one can better assess the role of outcome probability as intrinsic to the biasing effects in question, having eliminated alternative elements of computational difficulty and ambiguity. In doing so, it is possible to therefore answer whether probabilistic ambiguity alone is sufficient to sustain previously found effects. The change to binary outcomes bares a closer parallel to more social implications of these effects, such as impression formation and stereotyping (Anderson, 1965), wherein evidence is often categorically present or absent (e.g., adjective present or absent/high or low).

The final advantage of extending these effects into the health domain is the necessary implication for belief manipulation generalizability. By altering the context in which evidence is integrated, the belief itself also necessarily needs to be adapted to fit the new context. The consequent change in the content of the belief not only speaks further to the generalizability and robustness of the effects in question, but further allows for initial inferences in the degree to which beliefs are processed.

Accordingly, the methods below are a direct extension of Experiment 2, with the following changes:

7.1. Method

The context for the task was changed from a lottery task in which participants were required to assess the relative strength of two lottery machine algorithms, to a health domain in which the participant played the role of a physician prescribing medicines to patients. In this way, each trial was a new patient, presenting with the anonymised disease "Q". Participants were tasked with assessing the overall efficacy (i.e. across different patients) of two new medicines, anonymized to "K" and "Z". The cover story stated that each patient varied by genotype, and such variance may result in different responses to the two medicines. Such a change in context also required the comment section manipulation to change to reflect the health domain, so instead of a manipulation comment of "Machine A seemed luckier to me" as used in the lottery Experiments, comments instead reflected the medicine options (e.g., "I think medicine Z was the most effective" and "medicine Z was better than K."). In this way, communicated beliefs regarding the options still reflected the unquantified, directional hypotheses used previously.

As mentioned above, this change in context also allowed for the response format to change to a stricter, binary set of outcomes (instead of variable numbers of ball matches). A successful trial was defined as when the selected medicine "Cured" the disease (with the participant winning 3 points as a reward), and an unsuccessful trial as when the selected medicine had "No Effect" (costing the participant 1 point). As in previous experiments, participants could see the counterfactual outcomes for each trial (what the outcome for the patient would have been had they selected the other medicine), and were incentivized with increasing monetary bonuses for each 50-point boundary they crossed in earnings throughout the task, as in Experiments 1 and 2. Similarly, both the number of trial pre- and post-reversal remained the same, with 100 each side, and the probabilities of each option were again 70/30 pre-reversal, and 30/70 post-reversal.

Finally, in the demographics and questionnaire section following the main task and posteriors, Locus of Control and Revised Paranormal Belief Scale measures, which had previously failed to yield any relationships to both behavioural and judgement data, were replaced by an abbreviated Need for Closure scale (Roets & Van Hiel, 2011). Given prior literature associating Need for Closure with the propensity towards engagement, deliberation and entertainment of alternative hypotheses (Webster & Kruglanski, 1994) in individuals, this was hypothesised to have a potential impact on the biasing effects under investigation.

7.1.1. Hypothesis

The hypotheses for Experiment 3 are to replicate the contrast effects found in Experiment 1 and 2, extending these effects into a health context. Those who receive a belief that is initially supported (BiS) will choose the medicine indicated by the belief significantly more than controls, whilst those in the group that receive initially undermining evidence (BiU) will not show such a bias and be no different from the control group. These effects are also hypothesised to extend to posterior probability estimates,[9] replicating Experiment 2.

7.2. Results

7.2.1. Descriptives and processing

Based on the contrast code analysis of Experiment 2, a power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 3. Converting partial eta squared values for the contrast code analyses for the three dependent variables into Cohen's d (Cohen, 1992) effect sizes, using the smallest of these effect sizes, to detect a significant effect at the .05 level, with 80% power resulted in an average estimated sample size of 80 per group. Following the same procedure as Experiment 1, the groups sample sizes were increased by 33% to compensate for those failing manipulation checks, calculated from the failure rates of Experiment 1, resulting in a total N of 330. Given the changes to the paradigm, this was conservatively increased by 10% to 360.

Participants were recruited online using MTurk. Those who had

[9] In accordance with the context change from lotteries to health, the posterior probability estimate question changed to "What is the distribution of cures between the two medicines?", with a sliding scale from 100% K, though 50/50, to 100% Z.

taken part in the previous experiment were ruled out from participating. The 360 participants recruited were US based, randomized into either the BiS (103), BiU (119) or control (138) conditions. The average age was 35.52 years (SD = 11.181) and the sample was 47.2% female.

After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comment (if one had been presented in their condition), they were removed from subsequent analysis,[10] leaving 77 in the BiS group and 93 in the BiU group. The decision to remove those who failed was taken following the same protocol and reasoning as Experiment 1.

As expected, the change from showing one comment to several comments decreased the drop-out rate, although these differences (both between groups and between studies) were not significant. The following analyses were conducted using the remaining 299 participants, with an average age of 35.37 years (SD = 10.903) and 49.5% female.

7.2.2. Correlates

Need for closure did not correlate or interact with any of the effects under investigation, but was found to have an impact on the speed of learning across epochs in the first phase. Accordingly, Need for Closure was assessed in the convergence analysis. Age and gender variables were not correlated with any of the dependent variables.

7.2.3. Choice data

As can be seen in Fig. 4, the reversal point was harder to detect for all participants, but nevertheless by the point of reversal all groups had learnt a preference for the dominant medicine, and subsequently moved in the correct direction upon reversal. The proportion of choices in favour of the initially dominant medicine, both overall, and broken down into pre- (phase 1) and post-reversal (phase 2) proportions were the key variables of interest once again for running the contrast code analysis.

To assess whether Experiment 3 replicated the key findings of Experiment 2, the same contrast code analysis procedure was conducted for the three conditions (BiS, Control, BiU: 2, −1, −1), along with pairwise comparisons between groups.

The contrast code analysis of condition on the overall proportion of choices made was significant, $F(1, 296) = 8.462$, $p = 0.004$, $\eta^2 = 0.028$. Along with corroborating pairwise ANOVA analyses which show the BiS group was significantly higher than controls, $F(1, 205) = 4.859$, $p = 0.029$, $\eta^2 = 0.023$, and BiU, $F(1, 169) = 10.189$, $p = 0.002$, $\eta^2 = 0.058$, groups, whilst the difference between controls and BiU was not significant ($p = 0.407$).

Breaking this down by phase, the proportion of choices in the first phase (grey lines in Fig. 4, prior to reversal point), the contrast effect was significant, $F(1, 296) = 7.441$, $p < 0.001$, $\eta^2 = 0.025$. This was corroborated by pairwise ANOVA showing the BiS group was significantly higher than controls, $F(1, 205) = 7.044$, $p = 0.009$, $\eta^2 = 0.034$, and BiU, $F(1, 169) = 7.337$, $p = 0.007$, $\eta^2 = 0.042$, groups, whilst the difference between controls and BiU was not significant ($p = 0.975$).

These effects were not continued into the second phase (grey lines in Fig. 4, post reversal point), as the contrast effect was not significant, $p = 0.095$. This was explained by a pairwise ANOVA showing the BiS group was not significantly different from the control group ($p = 0.37$). Whilst the difference between controls and BiU was not significant ($p = 0.293$), the difference between the BiU and BiS groups was however significant in phase 2, $F(1, 169) = 4.73$, $p = 0.031$, $\eta^2 = 0.028$.

[10] Following the same protocol as Experiment 2, a significant effect was found for both the contrast code analysis of the overall choices, $F(1, 356) = 5.254$, $p = 0.021$, $\eta^2 = 0.015$, and posterior probability estimates, $F(1, 356) = 4.941$, $p = 0.027$, $\eta^2 = 0.014$, regardless of manipulation check removal.

7.2.4. Phase 1 convergence

Following the analysis protocol of Experiment 2, a repeated measures ANOVA was conducted to assess the degree of convergence in choice proportions over the course of phase 1, controlling for Need for Closure. The contrast code was the between-subjects grouping factor, and 10 trial epochs were within-subjects. The consequent interaction between epoch and grouping was significant, $F(2, 295) = 15.471$, $p < 0.001$, $\eta^2 = 0.095$, indicating a convergence of the contrast effect over the course of the phase.[11]

7.2.5. Posteriors

The posterior measure of estimated probability distribution (see Fig. 5) showed a significant contrast effect of condition, as found in the above behavioural measures, $F(1, 296) = 6.641$, $p = 0.01$, $\eta^2 = 0.022$.[12]

This was corroborated by pairwise ANOVA showing the BiS group was significantly higher than controls, $F(1, 205) = 4.3$, $p = 0.039$, $\eta^2 = 0.021$, and BiU, $F(1, 169) = 7.597$, $p = 0.006$, $\eta^2 = 0.044$, groups, whilst the difference between controls and BiU was not significant ($p = 0.272$).

A further series of t-tests revealed a primacy effect, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task, in the BiS, $t(76) = 4.788$, $p < 0.001$, 95% CI [4.27, 10.35], and control, $t(128) = 2.43$, $p = .0016$, 95% CI [0.61, 5.96], groups, but not the BiU group ($p = 0.533$).

Participant's binary preference posterior measure yielded a main effect of condition, $X2(1, N = 299) = 6.023$, $p = 0.049$, and showed strong primacy effects, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task, in BiS, $X2(1, N = 77) = 15.909$, $p < 0.001$, and control groups, $X2(1, N = 129) = 4.845$, $p = 0.028$, whilst BiU showed no such effect ($p = 0.351$). The confidence measure for the binary preference did not show any significant effects of condition, or order effects.

7.3. Discussion

The successful replication of the overall contrast effect found in both Experiments 1 and 2 despite the change of domain from gambling to health, speaks to the generalizability of the effects in question. The successful use of novel belief manipulation content, novel use of binary outcome types for experienced evidence, and the medicine prescribing context further validate the robustness of the biasing effects under investigation. However, the authors take pains to note that this is not a perfect replication. Although both overall and phase one choices reflect the patterns seen previously (including the convergence of choices across phase one), the re-emergence of the contrast effect does not quite reach significance in phase 2. It is however equally important to note that the contrast effect re-emerged in the posterior probability estimate, indicating the final judgement regarding the options is (just as in Experiment 2) a consequence of the interaction between belief and initial evidence.

One likely reason for this failure to replicate in phase 2, is the change from variable (and thus more computationally complex) outcomes in the gambling context, to the more obvious binary outcomes of

[11] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast effect in the first block in direct comparison to the assessment of the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first epoch, $BF_{10} = 9732.711$. In the final epoch, there was no longer support for the contrast effect, reaching the $< 1/3rd$ to be classified as strong support for the null, $BF_{10} = 0.265$. Both epoch analyses accounted for the effect of Need for Closure as mentioned in the correlates section.

[12] A Bayesian T-test was conducted to confirm the lack of a difference between controls and BiU was not due to insufficient power by looking for strong support of the null (a Bayes Factor of $< 1/3rd$). Substantial evidence was found in support of the null, $BF_{10} = 0.252$.

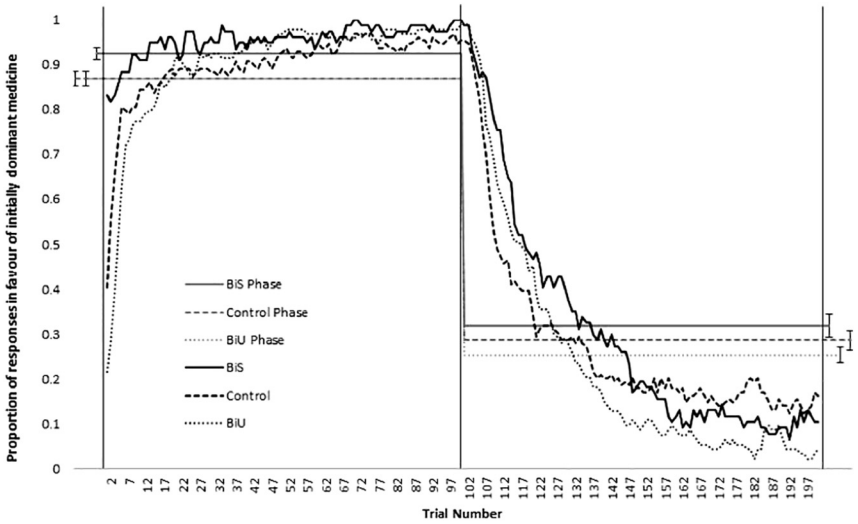


Fig. 4. Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially dominant medicine as participants move through the 200 trials (with reversal occurring at 100 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

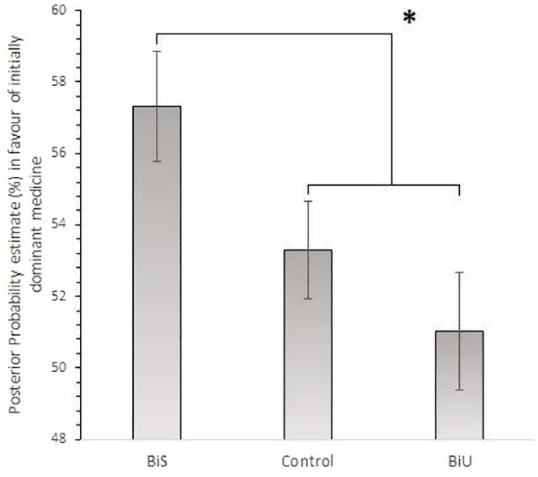


Fig. 5. Posterior probability estimates as a percentage of better outcomes. Greater than 50% reflects a preference for the initially dominant medicine, less than 50% indicates a preference for the initially suboptimal medicine. Outcomes split by group ($*p = 0.01$).

curing or failing to cure in the health context. This can be seen in visual inspection of the trend data shown in Fig. 4, wherein the point of convergence occurs earlier than in Experiment 2 (which is matched on both probability distribution and trial number), within the first half of the phase, rather than by the end of the phase. Further, the proportion of choices in all groups is substantially closer to maximisation (with the latter half of phase one showing all three group averages at about 90% optimal choices). Both of these trends are indicators that learning which option is optimal is easier, a situation that was present in Experiment 1 (which suffered similar ceiling effects likely due to unambiguous probability distributions), and successfully remedied via a subsequent increase in probabilistic ambiguity in Experiment 2. In other words, when first-hand learning is easier (re: less ambiguous evidence, whether through a reduction in computational complexity by going from variable to binary outcomes, as in Experiment 3, or through unambiguous probabilities, as in Experiment 1), the need for and use of communicated beliefs decreases (Ha & Hoch, 1989; Hoch & Ha, 1986).

Need for Closure was found to play a role in explaining some of the variance in choice behaviours, primarily in the first phase of the experiment. This finding fits with the Need for Closure literature and the role it plays in the length and degree of engagement with the process of consolidating on a single hypothesis (Webster & Kruglanski, 1994). To put this another way, those high in the need for closure more quickly came to a conclusion regarding which option to choose for the remainder of the phase. This finding provides an innovative demonstration of the impact individual differences in Need for Closure can have on extended evidence integration processes. It should further be noted that such effects were independent of manipulation condition, suggesting the consequent biasing effects run independently of this dimension.

8. Experiment 4

Experiment 4 set out to replicate the principal findings of Experiments 1, 2, and 3, whilst extending them into a reduced trial format. By reducing the number of trials in each phase, Experiment 4 sought to test the potential limits or thresholds for the consolidation/refutation effects on belief as consequence of the first phase in previous experiments. By seeking to extend such effects to instances where there is substantially less evidence to either refute or consolidate the communicated belief pre-reversal, it becomes possible to address questions regarding potential memory issues in the longer tasks contributing to previously found effects. Finally, such a change opens the avenue for future research into the limitations of previously found belief-initial evidence interactions.

8.1. Method

The design and procedure of Experiment 4 followed that of Experiment 3, wherein participants made choices between two medicines in an attempt to discern which is the more effective in curing a disease, some having been pre-exposed to a belief that indicates one of the options as superior. This belief is either supported (BiS) or undermined (BiU) by the first phase of evidence. The probability distribution of outcomes between the two options starts with one dominant (cures 70% of the time) and the other suboptimal (cures 30% of the time), before these probabilities then reverse halfway through the task. These choices are then followed by posterior measures assessing participants' end-of-sequence, overall judgements regarding the two options. Finally, these are followed by demographics and the Need for Closure scale.

The principle change in methodology from Experiment 3 to Experiment 4 was the total number of trial pre- and post-reversal were halved, resulting in two phases of 50 trials. Such a change also meant the pay scale for the task needed to be adjusted downwards to reflect a bonus scheme proportional to the reduced length of the task.

8.1.1. Hypothesis

The central hypothesis follows that of Experiments 2 and 3, namely a replication of the contrast effects found in both choice data and posterior judgements. To once again spell this out explicitly, those who receive a belief that is initially supported (BiS) will choose the medicine indicated by the belief significantly more than controls, whilst those in the group that receive initially undermining evidence (BiU) will not show such a bias and be no different from the control group. Accordingly, these effects are hypothesised to extend to this reduced trial format.

8.2. Results

8.2.1. Descriptives and processing

Based on a reduced phase contrast code analysis of Experiment 3, a power analysis was run using G*power (Faul et al., 2009, 2007) to estimate sample sizes required for Experiment 3. Converting partial eta squared values for the contrast code analyses for the three dependent variables into Cohen's d (Cohen, 1992) effect sizes, using the smallest of these effect sizes, to detect a significant effect at the .05 level, with 80% power resulted in an average estimated sample size of 80 per group. Following the same procedure as Experiment 1, the groups sample sizes were increased by 33% to compensate for those failing manipulation checks, calculated from the failure rates of Experiment 1, resulting in a total N of 270.

Participants were recruited online using MTurk. Those who had taken part in the previous experiment were ruled out from participating. The 270 participants recruited were US based, randomized into either the BiS (105), BiU (98) or control (68) conditions. The average age was 32.44 years (SD = 10.54) and the sample was 39.1% female.

After completing the task, all participants were asked a series of filter questions to determine whether the comment manipulation had been remembered. If participants had no recollection of a manipulation comment (if one had been presented in their condition), they were removed from subsequent analysis,[13] leaving 91 in the BiS group and 76 in the BiU group. The decision to remove those who failed was taken following the same protocol and reasoning as Experiment 1.

The following analyses were conducted using the remaining 229 participants, with an average age of 32.24 years (SD = 10.194) and 38.9% female.

8.2.2. Correlates

Need for closure, age and gender variables were not correlated with any of the dependent variables.

8.2.3. Choice data

As can be seen in Fig. 6, the reversal point was harder to detect for all participants, but nevertheless by the point of reversal all groups had learnt a preference for the dominant medicine, and subsequently moved in the correct direction upon reversal. The proportion of choices in favour of the initially dominant medicine, both overall, and broken down into pre- (phase 1) and post-reversal (phase 2) proportions were the key variables of interest once again for running the contrast code analysis.

To assess whether Experiment 4 replicated the key findings of Experiment 3, the same contrast code analysis procedure was conducted for the three conditions (BiS, Control, BiU: 2, −1, −1), along with pairwise comparisons between groups.

The contrast code analysis of condition on the overall proportion of choices made was significant, $F(1, 226) = 34.675$, $p < 0.001$, $\eta^2 = 0.133$. Along with corroborating pairwise ANOVA analyses which show the BiS group was significantly higher than controls, $F(1, 152) = 12.833$, $p < 0.001$, $\eta^2 = 0.078$, and BiU, $F(1, 166) = 46.12$, $p < 0.001$, $\eta^2 = 0.218$, groups, whilst the difference between controls and BiU was also significant to a lesser degree $F(1, 137) = 4.769$, $p = 0.031$, $\eta^2 = 0.034$.

Breaking this down by phase, the proportion of choices in the first phase (grey lines in Fig. 6, prior to reversal point), the contrast effect was significant, $F(1, 226) = 41.796$, $p < 0.001$, $\eta^2 = 0.156$. This was corroborated by pairwise ANOVA showing the BiS group was significantly higher than controls, $F(1, 152) = 20.682$, $p < 0.001$, $\eta^2 = 0.12$, and BiU, $F(1, 166) = 60.302$, $p < 0.001$, $\eta^2 = 0.268$, groups, whilst the difference between controls and BiU was not significant ($p = 0.073$).

These effects were continued into the second phase (grey lines in Fig. 6, post reversal point), as the contrast effect was again significant, $F(1, 226) = 10.354$, $p = 0.001$, $\eta^2 = 0.044$. This was explained by a pairwise ANOVA showing the BiS group was significantly higher than the BiU group $F(1, 166) = 13.688$, $p < 0.001$, $\eta^2 = 0.077$. However, the difference between controls and BiS, and controls and BiU groups were both not significant ($p = 0.078$, and .131 respectively).

8.2.4. Phase 1 convergence

Following the analysis protocol of Experiment 3, a repeated measures ANOVA was conducted to assess the degree of convergence in choice proportions over the course of phase 1. The contrast code was the between-subjects grouping factor, and the five 10-trial epochs were the within-subjects factor. The consequent interaction between epoch and grouping was significant, $F(2, 226) = 33.191$, $p < 0.001$, $\eta^2 = 0.227$, indicating a convergence of the contrast effect over the course of the phase.[14]

8.2.5. Posteriors

The posterior measure of estimated probability distribution (see Fig. 7) showed a significant contrast effect, as found in the above behavioural measures, $F(1, 226) = 8.460$, $p = 0.004$, $\eta^2 = 0.036$.[15]

Pairwise ANOVA showed the BiS was significantly higher than the BiU group, $F(1, 166) = 12.849$, $p < 0.001$, $\eta^2 = 0.072$. The differences between controls and BiS, and controls and BiU groups were not significant ($p = 0.167$ and 0.055 respectively).

A further series of t-tests revealed a primacy effect, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task, in the BiS, $t(76) = 4.788$, $p < 0.001$, 95% CI [0.814, 14.44], and control, $t(128) = 2.43$, $p = 0.016$, 95% CI [4.04, 11.64], groups, but not the BiU group ($p = 0.533$).

Participant's binary preference posterior measure yielded a main effect of condition, $X2(1, N = 229) = 10.626$, $p = 0.005$, and showed strong primacy effects, wherein the initial evidence favouring one medicine dominated the reversed evidence in the latter half of the task, in BiS, $X2(1, N = 91) = 40.89$, $p < 0.001$, control, $X2(1, N = 62) = 18.645$, $p < 0.001$, and BiU groups, $X2(1, N = 76) = 4.263$, $p = 0.039$.

The confidence measure for the binary preference showed a

[13] Following the same protocol as Experiment 2, a significant effect was found for both the contrast code analysis of the overall choices, $F(1, 268) = 24.114$, $p < 0.001$, $\eta^2 = 0.083$, and posterior probability estimates, $F(1, 268) = 8.024$, $p = 0.005$, $\eta^2 = 0.029$, regardless of manipulation check removal.

[14] This was further ratified by Bayesian ANOVAs conducted on both the first and last epochs, to assess the strength of support for the contrast effect in the first block in direct comparison to the assessment of the strength for the null in the final epoch. Decisive evidence was found for the contrast effect in the first epoch, $BF_{10} = 1.581*1011$. In the final epoch, there was no longer support for the contrast effect, but this did not reach the $< 1/3rd$ to be classified as strong support for the null, $BF_{10} = 2.058$. However, it should be noted that a Bayesian ANOVA on the subsequent epoch (first epoch post-reversal) indicated a reversal in trend from convergence to divergence upon reversal, as strong evidence is once again found for the contrast effect, $BF = 6.951$.

[15] A Bayesian T-test was conducted to assess the null effect between controls and BiU. Although there was a null effect, $BF_{10} = 0.997$, this did not reach the $< 1/3rd$ recommended Bayes Factor for strong support of the null.

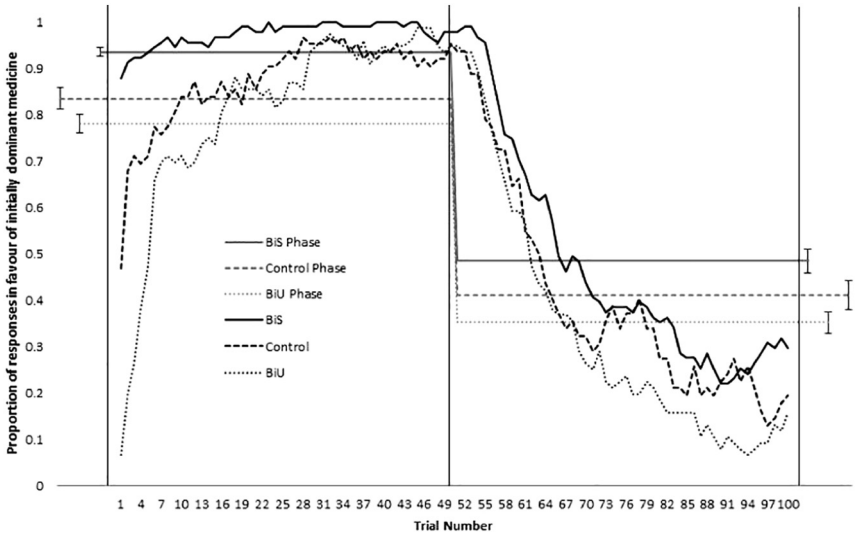


Fig. 6. Black lines show the within-group 7 trial windowed averages of proportion of choices made in favour of the initially dominant medicine as participants move through the 100 trials (with reversal occurring at 50 trial point), averaged on an individual level, then split into group averages. Grey lines show the averaged proportions of choices for each group, split by phase. Standard errors for phases are also shown.

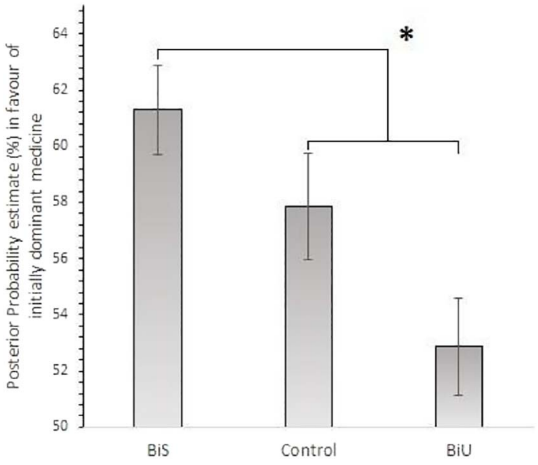


Fig. 7. Posterior probability estimate as a percentage of better outcomes. Greater than 50% reflects a preference for the initially dominant medicine, less than 50% indicates a preference for the initially suboptimal medicine. Outcomes split by group (*$p < 0.01$).

significant contrast effect, $F(1, 226) = 6.549$, $p = 0.011$, $\eta^2 = 0.028$, which was corroborated by pairwise analyses.

8.3. Discussion

The results of Experiment 4 replicate the effects found in Experiments, 1, 2, and 3, across all choice measures (overall, phase 1, and phase 2) and posterior probability estimates. This pattern of results demonstrates once again the importance of initial evidence in the validation of a communicated belief, and the subsequent integrative confirmation bias effects that occur as a consequence of consolidation. Furthermore, the replication of this pattern in Experiment 4 demonstrates that the threshold amount of evidence for consolidation or refutation of a communicated belief is lower than previous experiments have indicated. Such a finding raises questions regarding the limitations of these initial evidence effects for further research. In particular, the differences in biasing effects when reducing *amount* of evidence versus *clarity* of evidence.

However, as can be seen from visual inspection of Figs. 6 & 7, for the second phase choices and posteriors measures, although the contrast effect is significant, the difference between controls and the BiU group, although non-significant, does not show strong support for the null. Consequently, this finding suggests the potential proximity of the aforementioned threshold for belief refutation.

A further novel finding from Experiment 4 is the extension of the contrast effect to posterior confidence. Interestingly, this pattern demonstrates that those who show the most pronounced bias, i.e. the BiS group, also retain the highest degree of confidence in their [biased] choices and judgements.

9. General discussion

Through the use of on-line paradigms, participants played various lottery gambles and medical prescription tasks, choosing between two options, experiencing outcomes of both (chosen and counterfactual). Prior to this, participants were exposed to an on-line comment section that either contained entirely neutral comments (control group), or additionally contained a communicated belief regarding one of the two options being better. These beliefs were either initially supported (BiS) or undermined (BiU) by the subsequent probabilistic evidence (probabilities then reversed halfway through the task, rendering the options equally profitable overall). As such, the focus of this research has been to investigate the role of communicated beliefs in biasing the integration of evidence. Exploratory analysis from Experiment 1 found that a communicated belief required initially supporting evidence to procure a bias: When evidence changed, only those with a supported belief showed a re-emergence of confirmation bias even though evidence no longer favoured the belief (ruling out a normative “washing out” effect of a communicated prior, instead indicating an ongoing, active, biased integration). This reluctance to abandon the belief was not evident in participants who receive a belief that was initially undermined (ruling out conservatism or strong prior explanations). Importantly, all BiU group participants used in analyses could still *remember* the belief, and as such their similarity to controls was not a consequence of forgetting it. Further, the probabilistic reversal illustrates that all groups were actively learning throughout (see the sharp change in trend lines post-reversal in Fig. 1, Section 5.2). That is, as all groups remain sensitive to the change in probabilities, it is possible to rule out a differential in inattentiveness as an explanation of the bias found in the Belief initially Supported (BiS) group.

These effects were then replicated in Experiment 2, and further shown to extend beyond choice data to posterior judgements. This pattern of results was then replicated using a health context (Experiment 3) and with shorter phases (Experiment 4). Although both

binary preference and probability estimate posteriors showed strong overall primacy effects, in the latter case the interaction between belief and initial evidence lead to an exacerbation of the influence of initial evidence on the final judgements regarding the two options in the group for which the belief was initially supported. Furthermore, the design of the tasks makes it unlikely that biasing effects are driven by directional motivations to confirm the belief (e.g., demand characteristics). Moreover, the presence of counterfactual evidence assists in reducing selective exposure explanations.

The extension into the health domain (Experiment 3) highlights the robustness of the effects given changes to belief content, outcome types (variable to binary) and decision context (curing patients versus winning gambles). Such a finding strengthens the generalizability of the belief biasing mechanism proposed. Furthermore, the successful replication of the contrast (testing a difference between the Belief initially Supported (BIS) against the remaining groups) in the reduced trial number format of Experiment 4, speaks to the robustness of the effect. However, the pattern of results in Experiment 4 also raises some interesting questions regarding the limitations of a refutation period in belief-evidence interaction effects. Put another way, the finding in Experiment 4 that a linear contrast code is also a good fit for the pattern of results (i.e. there is starting to be a difference between controls and the Belief initially *Undermined* group) suggests the reduced amount of phase 1 evidence (100 down to 50 trials) is starting to impact belief refutation.

In contrast to most studies investigating confirmation bias effects, a novel manipulation was used in which additional motivations for confirmation (such as self-concept preservation and authority effects) were reduced. Such a reduction was due to the anonymous, on-line peer source of the directional beliefs regarding the two options, along with the seemingly incidental presentation of such information (no explicit instruction was given regarding the content of the comment section, and filter questions were used to identify and remove any participants that saw through this manipulation). Additionally, an accuracy incentive (performance-based pay) was used that was independent of (and antithetical to) belief-adherence motivations. Given the reduction of these traditional motivational explanations for confirmation bias, the fact an effect was found regardless points towards an alternative explanation.

9.1. Conclusions

Rather than motivational explanations, the current findings are better explained by an extension of cognitive explanations taken from research investigating self-generated hypotheses (Fischhoff & Beyth-Marom, 1983; Pyszczynski & Greenberg, 1987). The proposed mechanism behind the subsequent bias is that the communicated belief orients subsequent evidence as either confirming or contradicting it, in line with work on the constriction of the hypothesis space (Fischhoff & Beyth-Marom, 1983; Klayman, 1995; Nickerson, 1998). To explain further, those who receive a communicated belief start out with the hypothesis: “either this belief is true, or not”, whilst controls have no specific hypothesis to entertain. Such an assertion is supported by the naïveté of participants to the context in which evidence is evaluated, distinguishing this finding from work in which either evidence exposure (Harvey & Fischer, 1997; Yaniv, 2004; Yaniv & Milyavsky, 2007) or prior opinion (DeMarzo, Vayanos, & Zwiebel, 2003; Lord et al., 1979; Pitz, 1969) precedes second-hand information. This naïveté results in controls (i.e. those who receive no belief) starting with a neutral prior, whilst belief-recipients are informed (but not bound by) the communicated belief. In this way, evidence-exposure alone both forms a new hypothesis in controls, and validates or refutes communicated beliefs in recipients. Thus, subsequent deviations in learning can be attributed to the interaction of a communicated belief with the same evidence (as controls receive), rather than the additional (unaccounted for) interplay of prior beliefs and opinions which both communicated information and evidence integration will both be informed by (i.e. beliefs are not having to persuade the recipient away from a pre-existing informed prior).

Further, given the absence of cues to the credibility of the source (Briñol & Petty, 2009; Hahn et al., 2009), this hypothesis validation is, by inference, a validation of the source as credible or not. As such, initial evidence updates both of these likelihoods (the likelihood the belief is true, and in relation, the likelihood the source is credible), acting as a consolidation period. As such, those entertaining a hypothesis who experience initially undermining evidence for the belief not only consider the belief invalid, but the source (by proxy) as unreliable. In doing so, such individuals match the control group (i.e. a belief that has not been consolidated is equivalent to starting without a belief). Alternately, those who receive a belief that is confirmed by initial evidence would now have a consolidated hypothesis, and evaluate subsequent evidence in the forwarded integrative account of confirmation bias (Decker et al., 2015; Doll et al., 2009; Klayman, 1988, 1995; MacDougall, 1906; Staudinger & Büchel, 2013; Whitman et al., 2015). Such an evaluation stage of communicated knowledge in part fits within a propositional account of learning, in which the belief (or proposition) has its truth value either confirmed or refuted (Mitchell et al., 2009). However, given the interplay shown between accumulated evidence and communicated propositions, the effects are not readily explained by propositional or associative accounts in isolation.

The immediate refutation (or consolidation) of the communicated beliefs in phase one indicates a critical role of initial evidence, which rather than dominating or being dominated by the communicated belief, instead interacts with it. The presence of a bias in the BiS group (relative to the control group) and absence of bias in the BiU group indicates the belief must first be consolidated before biasing effects can occur beyond initial preferences. This role of initial evidence in either the consolidation or rejection of communicated beliefs has strong parallels to primacy effects found in other areas of psychology such as subjective probability learning (Peterson & DuCharme, 1967), causal judgements (Dennis & Ahn, 2001; Fugelsang & Thompson, 2003), and impression formation in social psychology (Anderson, 1965; Freund, Kruglanski, & Shpitzajzen, 1985; Kruglanski & Freund, 1983; Mann & Ferguson, 2015), where early evidence plays a critical role in the formation of opinion. However, this work differs from traditional primacy (Hogarth & Einhorn, 1992) in one key way: Although in the current experiments all experimental groups showed traditional primacy effects (as indicated by the overall proportions favouring the initially dominant option, as well as the posterior measures in Experiments 2, 3, and 4), those who received a communicated belief showed even greater susceptibility to early evidence (Staudinger & Büchel, 2013). This is demonstrated by the *re-emergence* of the bias in the initially supported belief condition (BiS) post-reversal, even after choices had converged prior to reversal. Further, at the point during the task of posterior measures, participants had seen the entirety of the evidence for both options (which were equal overall), yet the same pattern of bias was still found. This not only demonstrates the communicated belief had not been diluted by the first-hand evidence, but lends greater support to the potency of the interaction between the communicated belief and initial evidence, and the proposed consolidation, followed by integrative bias mechanism.

Once a belief had been consolidated, confirmation bias effects were found despite the presence of 2-sided evidence, mitigating selective evidence exposure as an explanation for confirmation bias effects (Doherty & Mynatt, 1990; Doherty, Mynatt, Tweney, & Schiavo, 1979; Klayman, 1995; Klayman & Ha, 1987). It is expected however, that in environments in which the individual can be selective in what they see (e.g., always selecting to take the medicine, so that the alternative of the disease disappearing without taking a medicine is never seen), the consolidation threshold would likely be crossed quicker (Blanco, Barberia, & Matute, 2014; Doherty et al., 1979; Yarritu,

Matute, & Vadillo, 2014). Similarly, it is expected that factors that increase participant's trust in the source of information (of which increasing the number of anonymous comments was one) will increase the degree of belief uptake (Yaniv & Kleinberger, 2000; Yaniv, 2004), along with factors that provide supplementary motivations for confirmation (Kunda, 1990).

9.2. Limitations

Despite the presentation of 2-sided evidence, in principle we cannot rule out a selective attention explanation, wherein people attended more to evidence that supported their belief, while missing the contradictory information. Such selective attention could therefore be considered a form of positive test strategy (Doherty et al., 1979; Klayman, 1988; Navarro & Perfors, 2011). Nevertheless, if a form of attentional selectivity to evidence were the mechanism at work in our current experiments, one might expect delayed learning post-reversal in the BiS group, whereby it would take a larger number of trials before choice proportions started moving away from their phase 1 choice proportions. Instead all groups rapidly moved away from their phase 1 proportion of choices upon reversal. Furthermore, similar delayed learning would be expected to amplify the effect of the prior in the BiU group during phase 1. To the contrary, we found a rapid negation of belief dependent on initial evidence. In other words, if participants were selectively attending to one of the options, negative outcomes for this option might in isolation be interpreted by the participant as short-term fluctuations (and thus not be worth switching away from). However, in the case of attending to both options, the negative outcomes in the selected option would be seen to co-occur with the counterfactual positive outcomes in the alternative, more rapidly indicating a change in the choice environment (resulting in a more immediate change of selected option), which is in line with the trends in choice data for all experiments. Although we cannot fully rule out selective attention in our current research, follow up lab studies could employ eye-tracking to assess what evidence participants are attending to, detecting possible asymmetries.

It is also important to note that effect sizes in the earlier experiments in the present paper are not large. However, it should be noted that as the methodology progresses through to latter experiments, effect sizes continue to increase. Such a trend is likely attributable to reducing sources of variance (e.g., from a gambling context that may invite more speculation and short-term fluctuations, to a more stable medical context with simpler outcomes, and from longer to shorter phases). Furthermore, given the purpose of the experiments presented within the present work has primarily been to make a theoretical point, and the paradigm itself is admittedly noisy, the small effect sizes may be an underestimation of real world instances of the biasing effects demonstrated. Further research is, however, needed to determine the extent of such a bias.

Two limitations pertaining to this work bear particular relevance for the extension of research investigating communicated belief effects. Firstly, when trying to ascertain whether an incidental belief has been noticed by participants, the use of a memory based manipulation check does carry a limitation: the criterion for passing applies greater scrutiny to the manipulation groups than the control group. In other words, an inattentive manipulation group participant would have been picked up by failing to remember the communicated belief, whilst those in the control group (who may be similarly inattentive), when answering that they had not seen anything, would have been marked as correct. This could result in an asymmetry in the number of (and quality) of participants in each group. However, as indicated within each experiment, when including all participants, the main effects did remain (just with greater variance), and further investigations on manipulation check failures found them no different from the control group. Importantly, as mentioned previously, this check (and removal of "forgetters") provided insight into the remaining similarity between controls and the BiU group used in analysis, showing this was not due to those in the BiU group forgetting the manipulation by the time the reversal had occurred.

Secondly, the use of MTurk has known shortcomings, notably the lack of experimental control (Goodman, Cryder, & Cheema, 2013), although several studies have shown strong replications of effects found in laboratory experiments using the site (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010). The advantage of such a method is its greater ecological validity, as participants not only show better representation of demographics relative to the majority of psychological study populations (Henrich, Heine, & Norenzayan, 2010). From the context of demand characteristics as an additional, unwanted motivation for belief confirmation, the fact participants did not see the manipulation as experimentally driven added real world applicability, and distinguishes this work from more artificial, lab-based assessments of integrative confirmation bias (Staudinger & Büchel, 2013). Further, there is a growing relevance in having a communicated belief manipulation that uses an on-line "comment section" in its natural setting, given such a context represents an ever-expanding and abundant source of communicated beliefs in the modern world.

9.3. Implications

Work in instruction effects has typically been lab-based, focusing on the altering of automatic processes via instruction (Doll et al., 2009; Mertens & De Houwer, 2016; Roswarski & Proctor, 2003; Van Dessel et al., 2015, 2016) through explicit experimenter influence. We argue that typical instruction influences are often rich in supplementary social explanations (such as demand characteristics). In this way, by highlighting and controlling such influences (e.g., through anonymous, rather than experimenter sources; placing accuracy incentives *against* belief-maintenance (instruction-based) motivations) we seek to add to this growing literature. Further, we seek to stress the critical impact of initial evidence in the adoption and maintenance of instruction-congruent behaviour.

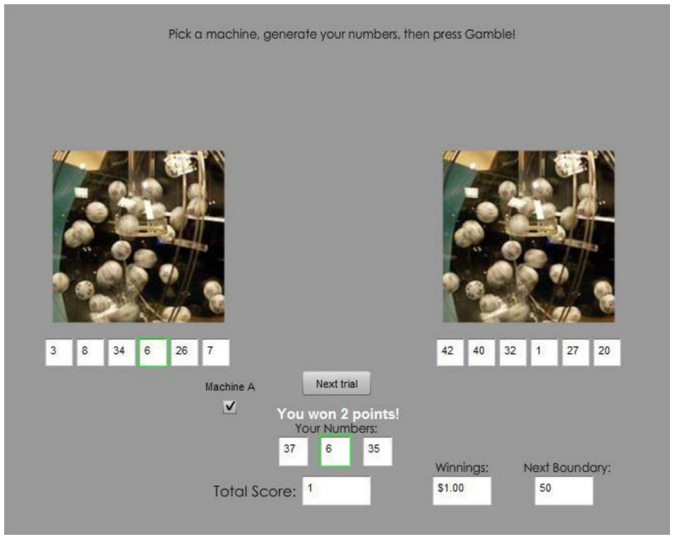
The instruction-based biasing effects discussed in this work have further implications in more applied domains. For example, work in stereotyping and impression formation has shown sensitivity to early evidence in the formation of negative assessments of character (Anderson, 1965). The effects found here suggest that the combination of a communicated belief (for example, a negative stereotype) and random fluctuations in early evidence (i.e. there will be a portion of recipients who happen to receive initial confirmation of this belief), results in an asymmetry that may help explain such beliefs prevalence despite their inaccuracy.

Similarly, research into placebo effects, for instance, relies on the patient's belief that a treatment will work, in spite of (unknown to the patient) the absence of a medically active ingredient. Typically, this research has investigated possible cues (setting, verbal suggestions), motivations (such as demand characteristics from doctor to patient) and "active agents" (such as caffeine) used to mimic a medical side effect (Wickramasekera, 1980) as highly impactful to the placebo effect (Wager & Atlas, 2015). The implication of the current research is to focus on providing initially supporting evidence to increase placebo efficacy. For example, by deploying treatments that start with a medically active component, initially supporting evidence (e.g., by decreasing symptoms) is provided for the belief of the patient that (s)he is in the treatment condition. Once the belief is consolidated the medically active component can be phased out of treatment in favour of a pure placebo, whilst retaining efficacy. Unfortunately, such initial sensitivity, when combined with short-term, random fluctuations, may also facilitate placebo responses in the case of various pseudo-scientific medical practices, such as homeopathy. Such effects also have ready application to other areas of research where there is an interplay of beliefs (or "priors") and evidence, including consumer research

(Ha & Hoch, 1989; Hoch & Ha, 1986), decision making (Nurek et al., 2014), causal learning (Yarritu, Matute, & Luque, 2015; Yu & Lagnado, 2012).

In summary, the present paper demonstrates that when the truth value of a communicated belief is uncertain, people use early experiences to validate the belief leading to an asymmetry in the *integration* of evidence. While the literatures on persuasion and learning from sequential evidence have evolved rather separately, the present research demonstrates that looking at the way communicated believes and experienced evidence interact may be a fruitful way of forwarding our knowledge of how (erroneous) action-outcome believes are formed.

Appendix A. Lottery experiment trial feedback screen



References

Abbott, K. R., & Sherratt, T. N. (2011). The evolution of superstition through optimal use of incomplete information. *Animal Behaviour*, *82*(1), 85–92. http://dx.doi.org/10.1016/j.anbehav.2011.04.002.

Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology*, *34*(1), 1–9. http://dx.doi.org/10.1037/h0021966.

Asch, S. E. (1955). Opinions and social pressure. *Scientific American*, *193*(5), 31–35. http://dx.doi.org/10.1038/scientificamerican1155-31.

Ayton, P., & Fischer, I. (2004). The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition*, *32*(8), 1369–1378.

Bar-Eli, M., Avugos, S., & Raab, M. (2006). Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, *7*(6), 525–553. http://dx.doi.org/10.1016/j.psychsport.2006.03.001.

Barron, G., & Leider, S. (2010). The role of experience in the gambler's fallacy. *Journal of Behavioral Decision Making*, *129*(October 2009), 117–129. http://dx.doi.org/10.1002/bdm.

Blanchard, S. J., Carlson, K. A., & Meloy, M. G. (2014). Biased predecisional processing of leading and nonleading alternatives. *Psychological Science*, *25*(3), 812–816. http://dx.doi.org/10.1177/0956797613512663.

Blanco, F., Barberia, I., & Matute, H. (2014). The lack of side effects of an ineffective treatment facilitates the development of a belief in its effectiveness. *PloS One*, *9*(1), http://dx.doi.org/10.1371/journal.pone.0084084.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. http://dx.doi.org/10.1016/j.obhdp.2006.07.001.

Bradbury, J. W., & Vehrencamp, S. L. (1998). Principles of animal communication. *Animal Behaviour*. http://dx.doi.org/10.1016/j.anbehav.2011.12.014.

Briñol, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, *20*(1), 49–96. http://dx.doi.org/10.1080/10463280802643640.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. http://dx.doi.org/10.1177/1745691610393980.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgement. *Journal of Personality and Social Psychology*, *66*(3), 460–473.

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology*, *55*, 591–621. http://dx.doi.org/10.1146/annurev.psych.55.090902.142015.

Cialdini, R. B., & Trost, M. R. (1998). Social influence: Social norms, conformity and compliance. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Vol. Eds.), *The Handbook of Social Psychology*(4th ed.). *Vol. 2*, (pp. 151–192). Boston: McGraw-Hill. http://dx.doi.org/10.2307/2654253.

Cimpian, A., Brandone, A. C., & Gelman, S.a. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, *34*(8), 1452–1482. http://dx.doi.org/10.1111/j.1551-6709.2010.01126.x.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*, 155–159.

Custers, R., & Aarts, H. (2010). The unconscious will: How the pursuit of goals operates outside of conscious awareness. *Science (New York, N.Y.)*, *329*(5987), 47–50. http://dx.doi.org/10.1126/science.1188595.

Dave, C., & Wolfe, K.. On confirmation bias and deviations from bayesian updating. (2003). Internet Access: http://www.Peel.Pitt.edu/esa2003/papers/wolfe_confirmationbias.pdf [Prieiga per Internetą 2011 02 24] .

Decker, J. H., Lourenco, F. S., Doll, B. B., & Hartley, C. A. (2015). Experiential reward learning outweighs instruction prior to adulthood. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 310–320. http://dx.doi.org/10.3758/s13415-014-0332-5.

DeKay, M. L., Miller, S. A., Schley, D. R., & Erford, B. M. (2014). Proleader and antitrailer information distortion and their effects on choice and postchoice memory. *Organizational Behavior and Human Decision Processes*, *125*(2), 134–150.

DeMarzo, P., Vayanos, D., & Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 909–968.

Dennis, M. J., & Ahn, W.-K. K. (2001). Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition*, *29*, 152–164. http://dx.doi.org/10.3758/BF03195749.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17. http://dx.doi.org/10.3389/fpsyg.2014.00781.

Doherty, M. E., & Mynatt, C. R. (1990). Inattention to P (H) and to P (D|~H): A converging operation. *Acta Psychologica*, *75*, 1–11.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*, 111–121.

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience*, *31*(16), 6188–6198. http://dx.doi.org/10.1523/JNEUROSCI.6486-10.2011.

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94. http://dx.doi.org/10.1016/j.brainres.2009.07.007.

Edwards, W. (1961). Probability learning in 1000 trials. *Journal of Experimental Psychology*, *62*(4), 385–394.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behaviour Research Methods*, *39*(2), 175–191.

Fiedler, K., & Freytag, P. (2004). Pseudocontingencies. *Journal of Personality and Social Psychology*, *87*(4), 453.

Fiedler, K., & Krueger, J. I. (2011). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgement and decision making*. New York: Taylor and Francis: Psychology Press.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*(3), 239–260. http://dx.doi.org/10.1037//0033-295X.90.3.239.

Freund, T., Kruglanski, A. W., & Shpitzajzen, A. (1985). The freezing and unfreezing of impressional primacy: Effects of the need for structure and the fear of invalidity. *Personality and Social Psychology Bulletin*, *11*, 479–487.

Frisch, K.v. (1950). *Bees: Their vision, chemical senses, and language*. Cornell University Press.

Frost, P., Casey, B., Griffin, K., Raymundo, L., Farrell, C., & Carrigan, R. (2015). The influence of confirmation bias on memory and source monitoring. *The Journal of General Psychology*, *142*(4), 238–252. http://dx.doi.org/10.1080/00221309.2015.1084987.

Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, *31*(5), 800–815.

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology*, *44*(6), 1110–1126.

Gilovich, T. (1993). *How we know what isn't so*. New York, NY: The Free Press.

Gilovich, T., Vallone, R., & Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, *17*, 295–314.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. http://dx.doi.org/10.1002/bdm.1753.

Goodwin, J. (2011). Accounting for the appeal to the authority of experts. *Argumentation*, *25*(3), 285–296. http://dx.doi.org/10.1007/s10503-011-9219-6.

Ha, Y., & Hoch, S. (1989). Ambiguity, processing strategy, and advertising-evidence interactions. *Journal of Consumer Research*, *16*(3), 354–360.

Hahn, U., & Harris, A. J. L. (2014). *What does it mean to be biased. Motivated reasoning and rationality. Psychology of learning and motivation — Advances in research and theory*(1st ed.). *61*Elsevier Inc.http://dx.doi.org/10.1016/B978-0-12-800283-4.00002-2.

Hahn, U., Harris, A. J. L., & Corner, A. (2009). Argument content and argument source: An exploration. *Informal Logic*, *29*(4), 337–367.

Hahn, U., Oaksford, M., & Harris, A. J. L. (2012). Testimony and argument: A Bayesian

perspective. In F. Zenker (Ed.), *Bayesian argumentation* (pp. 15–38). . http://dx.doi.org/10.1007/978-94-007-5357-0.
Harris, A. J. L., Hahn, U., Madsen, J. K., & Hsu, A. S. (2015). The appeal to expert opinion: Quantitative support for a Bayesian network approach. *Cognitive Science*, *39*(7), 1–38. http://dx.doi.org/10.1111/cogs.12276.
Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. http://dx.doi.org/10.1006/obhd.1997.2697.
Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61-83-135 http://dx.doi.org/10.1017/S0140525X0999152X.
Hertwig, R., & Ortmann, A. (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior*, *18*(1), 59–92. http://dx.doi.org/10.1080/10508420701712990.
Hoch, S., & Ha, Y. (1986). Consumer learning: Advertising and the ambiguity of product experience. *Journal of Consumer Research*, *13*(2), 221–233.
Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55. http://dx.doi.org/10.1016/0010-0285(92)90002-J.
Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2002). The Theory of Event Coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*(5), 849–878. http://dx.doi.org/10.1017/S0140525X01000103.
JASP Team (2016). *JASP*. Version 0.8.0.0.
Jessup, R. K., & O'Doherty, J. P. (2011). Human dorsal striatal activity during choice discriminates reinforcement learning behavior from the gambler's fallacy. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(17), 6296–6304. http://dx.doi.org/10.1523/JNEUROSCI.6421-10.2011.
Klayman, J. (1984). Learning from feedback in probabilistic environments. *Acta Psychologica*, *56*, 81–92.
Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 317–330.
Klayman, J. (1995). *Varieties of confirmation bias*. *32*, 385–418. http://dx.doi.org/10.1016/S0079-7421(08)60315-1.
Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*(2), 211–228. http://dx.doi.org/10.1037/0033-295X.94.2.211.
Klein, W. M., & Kunda, Z. (1989). Motivated person perception: Justifying desired conclusions. *Meeting of the Eastern Psychological Association* Boston.
Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, *19*(5), 448–468. http://dx.doi.org/10.1016/0022-1031(83)90022-7.
Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.
Leslie, S. J. (2008). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1), 1–47. http://dx.doi.org/10.1215/00318108-2007-023.
Levenson, H. (1973). Multidimensional locus of control in psychiatric patients. *Journal of Consulting and Clinical Psychology*, *41*(3), 397–404. http://dx.doi.org/10.1037/h0035357.
Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. http://dx.doi.org/10.1177/1529100612451018.
Liefooghe, B., De Houwer, J., & Wenke, D. (2013). Instruction-based response activation depends on task preparation. *Psychonomic Bulletin & Review*, *20*, 481–487.
Liefooghe, B., Wenke, D., & De Houwer, J. (2012). Instruction-based task-rule congruency effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1325–1335.
Lord, C., Ross, L., & Lepper, M. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109.
Ludvig, E. A., & Spetch, M. L. (2011). Of black swans and tossed coins: Is the description-experience gap in risky choice limited to rare events? *PloS One*, *6*(6), e20262. http://dx.doi.org/10.1371/journal.pone.0020262.
MacDougall, R. (1906). On secondary bias in objective judgments. *Psychological Review*, *13*(2), 97. http://dx.doi.org/10.1037/h0072010.
Mann, T. C., & Ferguson, M. J. (2015). Can we undo our first impressions?: The role of reinterpretation in reversing implicit evaluations. *Journal of Personality and Social Psychology*, *108*(6), 823–849. http://dx.doi.org/10.1037/pspa0000021.
Mertens, G., & De Houwer, J. (2016). Potentiation of the startle reflex is in line with contingency reversal instructions rather than the conditioning history. *Biological Psychology*, *113*(JANUARY), 91–99. http://dx.doi.org/10.1016/j.biopsycho.2015.11.014.
Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *The Behavioral and Brain Sciences*, *32*(2), 183-98-246 http://dx.doi.org/10.1017/S0140525X09000855.
Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134. http://dx.doi.org/10.1037/a0021110.
Newell, B. R., & Rakow, T. (2007). The role of experience in decisions from description. *Psychonomic Bulletin & Review*, *14*(6), 1133–1139. http://dx.doi.org/10.3758/BF03193102.
Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. http://dx.doi.org/10.1037//1089-2680.2.2.175.
Nurek, M., Kostopoulou, O., & Hagmayer, Y. (2014). Predecisional information distortion in physicians' diagnostic judgments: Strengthening a leading hypothesis or weakening its competitor? *Judgment and Decision making*, *9*(6), 572–585.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411–419. http://dx.doi.org/10.2139/ssrn.1626226.
Peterson, C. R., & DuCharme, W. M. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology*, *73*(1), 61–65. http://dx.doi.org/10.1037/h0024139.
Petty, R. E., & Cacioppo, J. T. (1984). Source factors and the elaboration likelihood model of persuasion. *Advances in Consumer Research*, *11*, 668–672.
Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*(3), 346–354. http://dx.doi.org/10.1037/h0023653.
Pitz, G. F. (1969). An inertia effect (resistance to change) in the revision of opinion. *Canadian Journal of Psychology/Revue Canadienne …*, *23*(1), 24–33. http://dx.doi.org/10.1037/h0082790.
Pitz, G. F., Downing, L., & Reinhold, H. (1967). Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *21*(5), 381–393. http://dx.doi.org/10.1037/h0082998.
Pohl, R. (2004). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Hove & New York: Psychology Press.
Priester, J. R., & Petty, R. E. (1995). Source attributions and persuasion: Perceived honesty as a determinant of message scrutiny. *Personality and Social Psychology Bulletin*, *21*, 637–654.
Pyszczynski, T., & Greenberg, J. (1987). Self-regulatory perseveration and the depressive self-focusing style: A self-awareness theory of reactive depression. *Psychological Bulletin*, *102*(1), 122–138. http://dx.doi.org/10.1037//0033-2909.102.1.122.
Rakow, T., & Newell, B. (2010). Degrees of uncertainty: An overview and framework for future research on experience-based choice. *Journal of Behavioral Decision Making*, *14*, 1–14. http://dx.doi.org/10.1002/bdm.
Roets, A., & Van Hiel, A. (2011). Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and Individual Differences*, *50*(1), 90–94. http://dx.doi.org/10.1016/j.paid.2010.09.004.
Roswarski, T. E., & Proctor, R. W. (2003). The role of instructions, practice, and stimulus-hand correspondence on the Simon effect. *Psychological Research*, *67*(1), 43–55. http://dx.doi.org/10.1007/s00426-002-0107-4.
Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social …*. Retrieved from http://psycnet.apa.org/journals/psp/50/4/703/.
Shanks, D. R. (2007). Associationism and cognition: Human contingency learning at 25. *Quarterly Journal of Experimental Psychology (2006)*, *60*(3), 291–309. http://dx.doi.org/10.1080/17470210601000581.
Siegrist, M., Cvetkovich, G., & Roth, C. (2000). Salient value similarity, social trust, and risk/benefit perception. *Risk Analysis*, *20*(3), 353–362. http://dx.doi.org/10.1111/0272-4332.203034.
Staudinger, M. R., & Büchel, C. (2013). How initial confirmatory experience potentiates the detrimental influence of bad advice. *NeuroImage*, *76*, 125–133. http://dx.doi.org/10.1016/j.neuroimage.2013.02.074.
Stern, H., & Cover, T. (1989). Maximum entropy and the lottery. *Journal of the American Statistical Association*, *84*(408), 980–985.
Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc. 8*(3), 220–247. http://dx.doi.org/10.1207/s15327957pspr0803.
Tobacyk, J., & Milford, G. (1983). Belief in paranormal phenomena: Assessment instrument development and implications for personality functioning. *Journal of Personality and Social Psychology*, *44*(5), 1029–1037. http://dx.doi.org/10.1037//0022-3514.44.5.1029.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110.
Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.
Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-based approach-avoidance effects: Changing stimulus evaluation via the mere instruction to approach or avoid stimuli. *Experimental Psychology*, *62*(3), 161–169. http://dx.doi.org/10.1027/1618-3169/a000282.
Van Dessel, P., Gawronski, B., Smith, C. T., & De Houwer, J. (2016). Mechanisms underlying approach-avoidance instruction effects on implicit evaluation: Results of a preregistered adversarial collaboration. *Journal of Experimental Social Psychology*. http://dx.doi.org/10.1016/j.jesp.2016.10.004.
Wager, T. D., & Atlas, L. Y. (2015). The neuroscience of placebo effects: Connecting context, learning and health. *Nature Publishing Group*, *16*(7), 403–418. http://dx.doi.org/10.1038/nrn3976.
Walton, D. (1997). *Appeal to expert opinion: Arguments from authority*. Penn State University Press.
Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. http://dx.doi.org/10.1080/17470216008416717.
Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*(6), 1049–1062. http://dx.doi.org/10.1037/0022-3514.67.6.1049.
Weiss-Cohen, L., Konstantinidis, E., Speekenbrink, M., & Harvey, N. (2016). Incorporating conflicting descriptions into decisions from experience. *Organizational Behavior and Human Decision Processes*, *135*, 55–69. http://dx.doi.org/10.1016/j.obhdp.2016.05.005.

Whitman, J. C., Takane, Y., Cheung, T. P. L., Moiseev, A., Ribary, U., Ward, L. M., & Woodward, T. S. (2015). Acceptance of evidence-supported hypotheses generates a stronger signal from an underlying functionally-connected network. *NeuroImage*, *127*, 215–226. http://dx.doi.org/10.1016/j.neuroimage.2015.12.011.

Wickramasekera, I. (1980). A conditioned response model of the placebo effect. *Biofeedback and Self-Regulation*, *5*(1), 5–18.

Wood, W. (2000). Attitude change: Persuasion and social influence. *Annual Review of Psychology*, *51*(1), 539–570. http://dx.doi.org/10.1146/annurev.psych.51.1.539.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. http://dx.doi.org/10.1016/j.obhdp.2003.08.002.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. http://dx.doi.org/10.1006/obhd.2000.2909.

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. http://dx.doi.org/10.1016/j.obhdp.2006.05.006.

Yarritu, I., Matute, H., & Vadillo, M. A. (2013). Illusion of control. *Experimental Psychology*, *32*(2), 38–47. http://dx.doi.org/10.1027/1618-3169/a000225.

Yarritu, I., Matute, H., & Vadillo, M. A. (2014). The Role of Personal Involvement. *Experimental Psychology*, *61*(1), 38–47.

Yarritu, I., Matute, H., & Luque, D. (2015). The dark side of cognitive illusions: When an illusory belief interferes with the acquisition of evidence-based knowledge. *British Journal of Psychology (London, England: 1953)*, *1–12*. http://dx.doi.org/10.1111/bjop.12119.

Yu, E. C., & Lagnado, D. A. (2012). The influence of initial beliefs on judgments of probability. *Frontiers in Psychology*, *3*(October), 381. http://dx.doi.org/10.3389/fpsyg.2012.00381.