Why and when beliefs change:

A multi-attribute value-based decision problem

Tali Sharot[1,2], Max Rollwage[2,3], Cass R. Sunstein[4] and Stephen M. Fleming[1,2,3]

[1]Department of Experimental Psychology, University College London, London WC1H 0AP, UK

[2]Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH

[3]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3AR, UK

[4]Harvard Law School, Harvard University, Cambridge, MA, USA

Abstract

Why people do or do not change their beliefs has been a long-standing puzzle. Sometimes people hold onto false beliefs despite ample contradictory evidence; sometimes they change their beliefs without sufficient reason. Here, we propose that the utility of a belief is derived from the potential outcomes associated with holding it. Outcomes can be internal (e.g., positive/negative feelings) or external (e.g., material gain/loss), and only some are dependent on belief accuracy. Belief change can then be understood as an economic transaction, in which the multidimensional utility of the old belief is compared against that of the new belief. Change will occur when potential outcomes alter across attributes, for example due to changing environments, or when certain outcomes are made more or less salient.

Key Words: Belief, Decision-making, Value, Confidence, Metacognition

In the current climate of increasing polarization, many people may assume that beliefs are rigid and fixed. Indeed, most individuals identify with the religious beliefs of their parents (Pew Research Centre, 2020) and by the age of seven many sport fans have established which teams they will support for the rest of their lives (Stephens-Davidowitz, 2017). Yet change happens. For example, in recent years many people have changed their beliefs regarding what constitutes workplace harassment and whether smoking in public venues is acceptable (Burns, 2014; Green Carmichael, 2017). Public health experts changed their minds on whether face masks can help reduce the spread of coronavirus, and on whether electric cigarettes are safe (Dutra et al., 2017; Greenhalgh et al., 2020). New information and experiences can and do lead people to change their beliefs.

John Maynard Keynes, the notable economist, is quoted as saying, "When I find new information I change my mind; What do you do?" (Clark, 1978) The answer, however, is not straightforward. Sometimes people do not alter their beliefs after receiving new information and other times they alter their views readily, with apparently little reason to do so (for review see Sharot & Garret, 2016). Such inconsistencies have baffled lay people as well as psychologists, economists and philosophers for decades (e.g, Kunda, 1990; Armor & Taylor, 2002; Moore & Small, 2008; Sunstein et al., 2016; Kappes et al., 2020; Klayman, & Ha, 1987).

Here, we propose that the value of beliefs (Benabou & Triole, 2016; Bromberg-Martin & Sharot 2020; Loewenstein, & Molnar, 2018) is composed of identifiable elements. Some of these elements are associated with the accuracy of a belief and some are not. By altering what they believe, people can gain or lose utility. Thus, the process of belief change can be understood as a (conscious or unconscious) process of weighing the value of an old belief against the expected value of a potential new belief. We show how such a conceptualization

can help explain why some beliefs seem intractable; why some beliefs change quickly; and why some strategies for promoting belief change succeed, while others fail dismally.

This perspective is not intended as a review of the literature on persuasion and/or influence (see Falk & Scholz, 2018 for a helpful review). Rather, our aim is to introduce the notion that the process of belief change can be understood as a multidimensional valuation problem. We suggest the process is analogous to multidimensional economic decisions. We marry recent findings from decision neuroscience (e.g., Blanchard, Hayden, Bromberg-Martin, 2015) with classic insights from psychology (e.g., Kunda, 1990) and behavioral economics (Loewenstein, 2006; Benabou & Triole, 2016) to describe the process.

Belief Change as a Multidimensional Valuation Problem

We conceptualize belief change as a *value-based decision*. The suggestion is that every belief carries a utility (Benabou & Triole, 2016; Bromberg-Martin & Sharot 2020; Loewenstein & Molnar, 2018). People will be more likely to change their beliefs when the expected utility of a new belief is greater than that of an old belief. The utility of a belief is derived by a summation of quantities along multiple dimensions. These dimensions can be roughly categorized into two groups: external outcomes of holding a belief and internal outcomes of holding a belief. The outcomes of holding a belief can be accuracy-dependent or accuracy-independent.

1. **External Outcomes**

   (i) **Accuracy-independent:** These refer to the external consequences of holding a belief, such as monetary rewards or social acceptance (Van Bavel et al., 2019), that are independent of whether the belief is accurate. For example, in certain societies people are more likely to find a job (positive external outcome) if they

hold certain religious views. These external outcomes (positive or negative) are independent of whether the belief itself is true or false.

(ii) **Accuracy-dependent:** These outcomes refer to the external benefits (or rewards) associated with holding an accurate belief and the costs (or punishments) associated with holding an inaccurate belief. For example, if people believe that the stock market will rise and invest in the market, they can gain money if they are correct, but will lose if they are incorrect. However, if they do not have any money to invest in the stock market (and are not advising others), the accuracy-dependent external outcomes are zero in this case. While many beliefs have direct accuracy-dependent consequences for the individual, as they guide actions with positive or negative consequences (e.g., believing whether cigarette smoking is good for you, whether coronavirus vaccines are safe, or whether a colleague is a friend or foe), many others do not (e.g., the positive or negative consequences of believing the earth is flat are not typically a function of the accuracy of the belief, unless the individual is navigating long distances, but instead involve social benefits of allying with like-minded others). Furthermore, other beliefs may not have a corresponding notion of accuracy at all, such as preferences ("chocolate ice cream is better than vanilla" or "dogs are better than cats") or beliefs about what is right and wrong ("people should not sacrifice animals for food").

2. Internal Outcomes

(iii) **Accuracy-independent**: Internal outcomes refer to the positive or negative cognitive and affective outcomes derived *directly* from a belief itself regardless

of whether there are external outcomes associated with the belief. These outcomes are often independent of whether the belief is accurate or not. For example, holding positive beliefs about oneself and the future (e.g., believing one will likely live a very long time or obtain a terrific job) can lead to a positive mental state (Charpentier et al., 2016; Loewenstein, 2006). This is because people are forward-looking agents who care about their future states (Brunnermeier & Parker's, 2005). A belief that a future state will be desirable leads to a current positive state known as 'positive anticipation' (Brunnermeier & Parker's 2005). Yet another example is holding beliefs with high certainty, which gives people a comforting sense that they understand the world around them.

(iv) **Accuracy-dependent**: Internal outcomes can also be accuracy-dependent. For example, holding a belief that one is likely to obtain a good grade can lead to (accuracy-independent) positive feelings in the present moment, but to great disappointment later when a failing grade is revealed (Rutledge et al., 2014). The latter is an internal outcome that is accuracy-dependent yet derived directly from the belief. That is, if one was to expect a failing grade the magnitude of disappointment would be negligible.

Internal and external outcomes can interact. For example, believing that one is likely to perform well on a job interview can in turn improve actual performance in the interview, increasing the likelihood of obtaining the job (Bandura, 1977; Benabu & Triole, 2002). However, exaggerated self-confidence can also be self-defeating, for example leading to the pursuit of sub-optimal endeavors (thus resulting in negative

external outcomes), but still maintained due to hedonic motives (e.g., internal outcomes) (Benabu & Triole, 2002).

We propose that expectations about all these different outcomes are implicitly combined to derive the overall utility of each belief . Forming a belief can thus be conceptualized as a multi-attribute value-based decision problem in which the aim is to hold a belief that has the highest value (most likely to lead to desirable outcomes), rather than necessarily forming the most accurate belief. People may incorporate these dimensions at an unconscious level (i.e., they do not necessarily have explicit access to these calculations). This is not unusual; the brain engages in many unconscious calculations that drive decisions (for example, estimating the speed and distance of an upcoming car before crossing the street) (Goschke, ; 1997; Pessiglione et al., 2008). Thus, while the brain may code for the value of the belief and estimate different outcomes, individuals will not necessarily have conscious access to this process and/or to the values of each attribute. This process may lead people to believe that the view with highest utility is the accurate one due to rationalization. Such a belief will feel subjectively justified, due to the automaticity of the belief-formation process (Festinger, 1962; Sharot, et al., 2010). When new evidence comes to light the difference in utilities of a potential new belief and old belief are compared. If the utility of a new belief is greater than an old belief, then a change in belief is likely.

This framework can account for cases in which people do not change their beliefs in the face of highly credible new evidence. For example, individuals fail to adequately alter their beliefs in the face of information that points towards unpleasant conclusions, such as learning that the likelihood of an adverse event (such as an accident or illness) is worse-than-expected (Sharot et al., 2011; Kappes, & Sharot, T. 2019; Moutsiana et al., 2015), learning that others view them

as less attractive then they thought (Eil & Rao, 2011), learning that they are likely to earn less than they expected (Mobius et al., 2011), or learning their preferred presidential candidate is lagging behind in the polls (Tappin et al., 2017). In all these cases individuals may hold onto inaccurate beliefs that are associated with non-accuracy-dependent outcomes (e.g., the positive feeling of maintaining a belief that it is pleasant to have) that are greater than the external accuracy-dependent outcomes.

A similar pattern of belief updating has been observed in reinforcement learning tasks, where participants are required to learn which of two cues is associated with the greatest reward. A larger learning rate is observed in response to unexpected positive outcomes than to negative outcomes (Lefebvre et al., 2017). Interestingly, the bias is observed only when participants select between cues themselves (that is when they have control over the outcomes) and not when a computer makes the choices for them (Chambon et al., 2020). In other words, participants are amplifying the belief that their choices are correct – a belief which is internally rewarding. It has been suggested that such a learning pattern (aka "choice-confirmation bias") can also lead to greater external rewards in some situations (Chambon et al., 2020). In such cases the resulting belief will have high value due to both internal and external outcomes. Note, however, that in different contexts a positive bias in belief updating has been observed even in situations where people have no control over the outcomes, such as in updating beliefs regarding whether one carries the Huntington gene (Oster et al., 2013).

When a person's environment or situation changes the value of accuracy-dependent outcomes relative to non-accuracy-dependent outcomes can vary. In environments rife with threat, the external accuracy-dependent cost of underweighting negative information could be particularly high. For example, the outcome of holding on to a belief that one is immune to a deadly

infectious virus amid a global pandemic may be grave. Indeed, it has been shown that exposing participants to a threatening environment increases the likelihood that they will adequately change their beliefs in response to unpleasant information (Garrett et al., 2018; Globig et al., 2021).

Or consider an individual who grows up in an environment where social acceptance is conditional on holding conservative beliefs, but who then moves to a town where both conservatives and liberals are socially accepted. The external non-accuracy dependent outcomes of holding conservative beliefs are reduced or eliminated, and hence the individual may shift their beliefs based on the other dimensions. In other words, people may change their beliefs when their environment changes, because those changes bring with them alterations to the value of the different dimensions of a belief. Because people experience different environments and have different personalities and values (for example, some people may care more/less about social acceptance) the utility of a belief will be different for different people, which can lead to diverse beliefs within a population. Political polarization might well resolve from this process, as when one environment rewards a certain set of beliefs, and another environment rewards a different set of beliefs; a "belief subsidy" is some places turns into a "belief tax" in others.

Belief change can also occur when one of the above attributes is made more salient. For example, if a person is nudged to consider accuracy, they may give more weight to accuracy-dependent outcomes than they would otherwise and consequently shift their beliefs. Indeed, a study reported that priming subjects to consider the veracity of social media posts by asking them to rate the accuracy of a single post subsequently resulted in reduced sharing of other

false information (Pennycook et al., 2021). However, whether this manipulation also reduced the likelihood that subjects believe these posts to be true was not tested.

Just as the valuation and comparison of material goods can involve biases and heuristics (Tversky, A., & Kahneman, 1974) so could people's assessment of the value of beliefs, which could lead to mistaken judgments about the benefits and costs of changing beliefs. For example, people might overestimate the short-term adverse emotional impact of a new belief (about personal vulnerability to some health risk for example), partially because they underestimate their ability to adapt to negative information (Sharot & Sunstein, 2020). In some cases, people might hold onto their beliefs more tenaciously than they should, given the expected value of changing them, and in other cases, they might change their beliefs too readily, given that same expected value.

The Role of Confidence and Metacognition in Belief Change

The multidimensional framework described above is analogous to other multidimensional economic decision problems (for review Busemeyer et al., 2019). For instance, determining the subjective value of a banana is a multi-dimensional estimation problem. An agent needs to estimate how tasty the banana will be, how much sugar and fiber it has, the current level of sugar and fiber in one's body, and so forth (Maier et al., 2020). In turn, people may have different levels of uncertainty around their estimate related to each dimension (*dimensional uncertainty*). For example, you may be very certain about how tasty the banana will be but not about the amount of sugar in it. Conversely, a person may be unsure about whether holding religious beliefs will facilitate or impede job security or how they may feel if they no longer held such beliefs. Uncertainty about a dimension will usually reduce the impact of this dimension on the overall utility calculation, in line with Bayesian rules of information

integration where more precise signals are weighted more heavily (Ernst & Banks, 2002). To exemplify this point, scientists might have high certainty in the effects of vaccinations on COVID-19 case numbers (external accuracy-dependent dimension) as they are educated in the scientific method, leading to a larger influence of this dimension on their overall belief. In comparison, people who are less familiar with the scientific method might feel less certain about the accuracy-dependent dimension, and thus, show less influence of this dimension on their overall belief. Moreover, uncertainty about any of the dimensions will contribute to low certainty in the overall integrated value of a belief– *belief uncertainty*.

Uncertainty about a belief or value is conceptually distinct from confidence in a decision about which belief to hold. This is analogous to the distinction drawn between confidence and certainty in other realms of decision-making (Pouget et al., 2016). For instance, when choosing between material options (such as between a banana and a dragon fruit) you may have higher certainty in the value of a banana than the value of a dragon fruit, with which you might have less experience. Choosing between the options then gives rise to different degrees of *decision confidence,* a quantity that is thought to be related to the difference in the distributions of the value of one option (e.g., banana) over another option (e.g., dragon fruit) (De Martino et al., 2013). The width of each value distribution is inversely proportional to the certainty about the value, which in turn affects how much the distributions overlap. When the value distributions are overlapping, deciding between the options is typically hard, and decision confidence is typically low. When the distributions are well-separated, the decision is easy and confidence is high (De Martino et al., 2013). Similarly, when deciding whether to change one's belief, a value comparison between opposing beliefs can be made and the greater the distance between the two value distributions, the greater the confidence in the adopted belief. For instance, while a person may be unsure about the overall utility of being an atheist (high *belief uncertainty*),

they could still hold high *decision confidence* that for them it is preferable to following Pastafarianism, due to a clear relative difference in value.

In standard decision-making tasks, people are more likely to gather additional information when 'decision confidence' is low (Desender et al., 2018b, 2019; Meyniel, Schlunegger, et al., 2015; Schulz et al., 2020, Folke et al., 2016, De Martino et al., 2013, Fleming et al., 2018). The process should be similar for beliefs; if people are not confident in an initial belief that vaccines are ineffective and unsafe, for example, they might continue to ask for new information, eventually changing their belief. In that sense, highlighting what people do not know may be an effective way to trigger information-seeking. Information gathering can take various forms; actively seeking new information (e.g., looking up studies on vaccine efficacy) (Desender et al., 2018b, 2019; Schulz et al., 2020; Gershman, 2018; E. Schulz et al., 2019; E. Schulz & Gershman, 2019), resampling of internal evidence (e.g., recalling a past conversation with one's physician about vaccine efficacy; Lee & Daunizeau, 2020), or paying attention to information accidently encountered in the media environment (Hornik & Niederdeppe, 2008). If, however, the potential outcomes of a belief (internal or external) are negligible, people are unlikely to invest time and effort in seeking information (for example, one may be highly uncertain whether vaccines are safe, but will not bother to investigate the matter if they expect never to have access to vaccines).

Besides effects on information seeking, low decision confidence has itself been found to make it more likely that new evidence will induce belief change regardless of whether the new information was actively sought out or not (Meyniel, 2020; Rollwage et al., 2020). Confidence levels can thus be adaptive in optimally allocating resources towards acquiring and processing valuable information (Lee & Daunizeau, 2020; Meyniel, 2020). In this sense, confidence plays

the role of an internal control mechanism indicating the need (or no need) for further processing and adapting the receptiveness to new information accordingly. As suggested above, (high) confidence may itself be a component of an accuracy-independent internal outcome, in that a comforting feeling of confidence in the world may itself be intrinsically valuable. This is in keeping with studies that have shown value and confidence signals are both represented in a similar region of ventromedial prefrontal cortex (De Martino et al., 2013; Lebreton et al., 2015) and with the demonstration of interactions between monetary incentives and confidence (Lebreton et al., 2018).

How useful these control signals are will depend on their alignment with the true underlying distribution of the belief utilities (Rollwage & Fleming, 2021; Schulz et al., 2021). Previous work has shown that confidence can be influenced by factors extraneous to the decision (for instance, fluency and arousal). If confidence is poorly aligned with the true underlying distributions, people might be confident even though they should not be, which would lead them not to invest mental effort in changing beliefs even when there is considerable belief uncertainty. Conversely, people might feel uncertain even though they should not, which could drive them towards a suboptimal belief change.

How well (decision) confidence aligns with true performance is known as metacognitive ability (Fleming et al., 2012; Fleming & Lau, 2014). People with high metacognitive ability will be very confident in their decisions when they are correct and not so confident when they are incorrect. Metacognitive ability is typically measured with respect to judgements that have a ground truth, such as the accuracy of a perceptual decision (e.g., ‘is an array of dots moving right or left?’ ‘how confident are you?’). But the notion of metacognition can also be extended to belief utility. When metacognitive ability is high, people will tend to have high confidence

in high utility beliefs and low confidence in suboptimal beliefs, motivating them to invest mental effort to potentially change their beliefs in the latter case. It is thus possible that increasing people's metacognitive abilities, for example through training (Carpenter et al., 2019), could increase openness to new information specifically is cases when it could be helpful for ensuring beliefs and values align.

Policy Applications

We have suggested that the value of a belief can be understood as a weighted summation of four types of belief outcomes that follow a 2 × 2 categorization (external and internal outcomes that are either dependent or independent on accuracy). Policymakers and practitioners may find it useful to consider *all* four "boxes" when attempting to predict and/or alter people's beliefs.

Many policies requiring disclosure of information are designed to alter beliefs. For example, information regarding health and safety, or labels informing consumers about fuel economy, are meant to bring consumers' beliefs into accordance with reality. Regulators often assume that consumers, workers, investors, and others care only about what is accurate, which means that if they are presented with the truth, they will believe it so long as it is credible (Food & Drug Administration, 2011). For reasons sketched above, that assumption might well be wrong. As we have seen, people also care about accuracy-independent dimensions of holding beliefs, including how beliefs make them feel.

The implication is that when policymakers (as well as advocates and marketers) are seeking to promote belief change (in the interest of health or safety, for example), they should also pay close attention to people's expectations about the internal outcomes of belief change (Sunstein, 2019) as well as perceptions of external outcomes that are not accuracy-dependent. If they do,

they might be able to recast or frame information in such a way as to make belief change more appealing.

As an example, consider the campaign to persuade people to believe in the safety and efficacy of COVID-19 vaccines. Most private and public institutions focus only on communicating data indicating the efficacy and safety of the vaccine (external accuracy-dependent outcome). Future studies should examine if highlighting accuracy-independent outcomes, such that learning that one is immune will greatly reduce anxiety (internal accuracy-independent outcome) or that people who believe in vaccine efficacy are more respected by their relevant peers (external accuracy-independent outcome), will increase beliefs in vaccine efficacy.

We further speculate that when fear-appeals work (Tannenbaum et al., 2015), it is because they can generate positive internal and external belief outcomes. Most fear appeals highlight a danger (COVID can lead to death) alongside a controllable solution (get vaccinated). Such fear appeals may be effective because the promoted belief ('vaccines work') has both positive accuracy-dependent external outcomes (people who hold such a belief will be more likely to get vaccinated and thus increase disease protection) and accuracy-independent internal outcomes (believing vaccines work reduces fear).

These points also bear on effective responses to misinformation and "fake news." In some cases, factual corrections do not work, in part because people do not want to believe them for reasons unrelated to accuracy (Van Bavel et al., 2020). In extreme cases, they can actually backfire, fortifying people's commitment to the belief that were supposed to be debunked (Nyhan et al., 2014). One reason may be people's judgment that if they changed their belief, they would in some sense suffer (perhaps because the new belief would endanger their

affiliation with generally like-minded others, perhaps because it would threaten their sense of identity, perhaps because it would make them feel sad or afraid). The implication is that if the correction can be made in a way that does not threaten people's affiliations or self-understanding, or the essentials of their view of the world, it is more likely to be effective (Kahan, 2017). "Surprising validators," who are not expected to endorse a new belief (such as a conservative politician who supports gay rights) but who are credible to those who are considering whether to do so, can succeed in promoting belief change in part for this reason (Glaeser, E., & Sunstein, 2014). If a new belief about (say) personal safety and health seems more like an opportunity rather than a threat, people may be more likely to be drawn to it.

The current framework may also hold intriguing implications for formal educational settings. The role of confidence and metacognition in guiding students' learning and study time has been long appreciated (Metcalfe, 2009). However, students also hold broader beliefs about their abilities (self-efficacy beliefs; Bandura, 1977) that impact upon future performance (Greven et al., 2009) and which may have both accuracy-dependent and accuracy-independent components.

Concluding Remarks

We suggest that a person's goal is to hold beliefs that carry maximum utility. The utility of a belief is equal to the weighted summation of the potential outcomes of holding that belief. Some of these potential outcomes are dependent on the accuracy of the belief, but some are not. For example, the outcomes of holding a religious belief may include reduced stress and social acceptance, neither of which are dependent on the accuracy of that belief. The outcomes of holding a belief about personal vulnerability to health risks may include fear and sadness, which people prefer to avoid.

It follows that the process of belief change is not necessarily an attempt to improve the accuracy of a belief, but rather to adopt a belief with higher utility. Sometimes belief change may not be observed even when highly credible new evidence, inconsistent with the current belief, is introduced; the accuracy-independent costs of changing one's beliefs might be perceived to be too high. Sometimes belief change may occur without any new evidence at all, but simply because the utility of holding it suddenly increases (for example due to a new environment, where external rewards are given to those who hold the new belief). Importantly, exposing individuals to new evidence to correct a false belief may not be sufficient for belief change in cases when a potential new belief does not carry higher utility than an old belief. This point underscores the importance of considering all relevant dimensions of a belief when aiming to elicit belief change.

Glossary

Belief: the acceptance that a proposition is true.

Belief utility: a quantity which reflects the benefit to oneself of accepting that a proposition is true.

Decision confidence – subjective feeling that a chosen course of action is optimal relative to others, often modeled as the probability that a decision is correct. In the case of belief, it is the subjective feeling that a belief has greater utility relative to alternative beliefs.

Belief uncertainty – uncertainty about the value of a belief.

Metacognition – the capacity to reflect on, monitor and control other cognitive states or processes.

Metacognitive ability – the extent to which confidence tracks performance, or distinguishes between correct and incorrect decisions (also known as metacognitive sensitivity).

Accuracy-dependent external outcomes of a belief: the external rewards (such a monetary gain) associated with holding an accurate belief and the punishments (such a monetary loss) associated with holding an inaccurate belief.

Accuracy-independent external outcomes of a belief: the external rewards or losses (such as social acceptance) of holding a belief that are independent of whether the belief is accurate.

Accuracy-independent internal outcomes of a belief: positive or negative cognitive and affective outcomes (such as feelings of joy, sadness, uncertainty) derived directly from holding a belief, regardless of whether the belief is true or false.

Accuracy-dependent internal outcomes of a belief: positive or negative cognitive and affective outcomes (such as feelings of joy, sadness, uncertainty) derived from holding a belief, which is contingent on whether the belief is true or false.

Outstanding Questions Box

A prediction arising from our framework is that the brain codes for the value of belief using similar neural architectures and computational rules as it does the value of material rewards and losses. The value of material goods is coded by the midbrain dopaminergic areas (e.g., the VTA and SN), the striatum and parts of the frontal cortex (e.g., the OFC). Does the same system code for the value of beliefs and is the neurotransmitter dopamine, which is central for processing the value of material rewards, also important for coding the value of beliefs?

If beliefs have value just like material goods, a prediction arising is that they are susceptible to the same biases and heuristics commonly observed in value-based decision making. To what extent is the value of belief context-dependent or subject to framing effects? For example, will the value of a belief alter when it is considered alongside other beliefs of high/low value?

Decision-making capabilities and cognitive flexibility are often assumed to be critical for deriving accurate beliefs. The conceptualization that people optimize for belief utility rather than accuracy makes the (counterintuitive) prediction: could greater decision-making capabilities and cognitive flexibility increase the likelihood of deriving inaccurate beliefs under certain circumstances (i.e. when accuracy-independent outcomes are especially pronounced)?

Does belief formation always proceed unconsciously, and feel subjectively justified? Or are people aware of the structure of their beliefs?

How are the utilities of competing beliefs compared? Is the overall utility of one belief compared to the other, is each dimension compared separately (Noguchi & Stewart, 2018), or are simple heuristics utilized?

Are the expected outcomes of a belief converted to a common currency, and if so how?

Can promoting (domain-general) metacognitive abilities facilitate belief change / openness to new information?

Acknowledgments

We thank Irene Cogliati Dezza, Bastien Blain, Moshe Glickman, Liron Rozenkrantz, Chris Kelly, Valentina Vellani, Laura Globig, Sarah Zheng, Gaia Molinaro and Christina Maher for comments on previous versions of the manuscript. TS is funded by a Wellcome Trust Senior Research Fellowship (214268/Z/18/Z). SMF is funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and a Philip Leverhulme Prize from the Leverhulme

Trust. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (206648/Z/17/Z). The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society.

References

1. Armor, D. A., & Taylor, S. E. (2002). When predictions fail: The dilemma of unrealistic optimism.

2. Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. Psychological review, 84(2), 191.

3. Bendixen, L. D. (2002). A process model of epistemic belief change.
4. Bénabou, R., & Tirole, J. (2002). Self-confidence and personal motivation. *The quarterly journal of economics*, *117*(3), 871-915.
5. Bénabou, R., & Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, *30*(3), 141-64.
6. Brunnermeier, M. K., & Parker, J. A. (2005). Optimal expectations. *American Economic Review*, *95*(4), 1092-1118.
7. Boldt, A., Blundell, C., & De Martino, B. (2019). Confidence modulates exploration and exploitation in value-based learning. Neuroscience of consciousness, 2019(1), niz004.

8. Blanchard, T. C., Hayden, B. Y., & Bromberg-Martin, E. S. (2015). Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. Neuron, 85(3), 602-614.

9. Burns, D. (2014). How far we have come in the last 50 years in smoking attitudes and actions. Annals of the American Thoracic Society, 11(2), 224-226.

10. Bromberg-Martin, E. S., & Sharot, T. (2020). The value of beliefs. Neuron, 106(4), 561-565.

11. Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. Trends in cognitive sciences, 23(3), 251-263.

12. Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019). Domain-general enhancements of metacognitive ability through adaptive training. Journal of Experimental Psychology: General, 148(1), 51.

13. Chambon, V., Théro, H., Vidal, M., Vandendriessche, H., Haggard, P., & Palminteri, S. (2020). Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, *4*(10), 1067-1079.

14. Charpentier, C. J., De Neve, J. E., Li, X., Roiser, J. P., & Sharot, T. (2016). Models of affective decision making: how do feelings predict choice?. Psychological science, 27(6), 763-775.

15. Clark L. H. Jr. (1978) Wall Street Journal.

16. De Martino, B., Fleming, S. M., Garrett, N., & Dolan, R. J. (2013). Confidence in value-based choice. Nature neuroscience, 16(1), 105-110.

17. Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. Psychological science, 29(5), 761-778.

18. Desender, K., Murphy, P., Boldt, A., Verguts, T., & Yeung, N. (2019). A postdecisional neural marker of confidence predicts Information-Seeking in Decision-Making. Journal of Neuroscience, 39(17), 3309-3319.

19. Dutra, L. M., Grana, R., & Glantz, S. A. (2017). Philip Morris research on precursors to the modern e-cigarette since 1990. Tobacc

20. Eil, D., & Rao, J. M. (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. American Economic Journal: Microeconomics, 3(2), 114-38.

21. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nat. 2002;415:429–433. doi: 10.1038/415429a.

22. Falk, E., & Scholz, C. (2018). Persuasion, influence, and value: Perspectives from communication and social neuroscience. Annual review of psychology, 69, 329-356.

23. Festinger, L. (1962). Cognitive dissonance. Scientific American, 207(4), 93-106.

24. Fleming, S. M., Van Der Putten, E. J., & Daw, N. D. (2018). Neural mediators of changes of mind about perceptual decisions. Nature neuroscience, 21(4), 617-624.

25. Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1594), 1338-1349.

26. Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. Frontiers in human neuroscience, 8, 443.

27. Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. Nature Human Behaviour, 1(1), 1-8.

28. Food & Drug Administration. Fed. Regist. 76, 36628 (2011)

29. Garrett, N., González-Garzón, A. M., Foulkes, L., Levita, L., & Sharot, T. (2018). Updating beliefs under perceived threat. Journal of Neuroscience, 38(36), 7901-7911.

30. Goschke, T. (1997). Implicit learning and unconscious knowledge: Mental representation, computational mechanisms, and brain structures.

31. Greenhalgh, T., Schmid, M. B., Czypionka, T., Bassler, D., & Gruer, L. (2020). Face masks for the public during the covid-19 crisis. Bmj, 369.

32. Green Carmichael, S. (2017) Have Our Attitudes About Sexual Harassment Really Changed? Harvard Business Review.
33. Greven, C. U., Harlaar, N., Kovas, Y., Chamorro-Premuzic, T., & Plomin, R. (2009). More than just IQ: School achievement is predicted by self-perceived abilities—But for genetic rather than environmental reasons. Psychological Science, 20(6), 753-762.
34. Globig L., Witte K., Feng G. & Sharot T. (in press) Under threat weaker evidence is required to reach undesirable conclusions. Journal of Neuroscience
35. Glaeser, E., & Sunstein, C. R. (2014). Does more speech correct falsehoods?. The Journal of Legal Studies, 43(1), 65-93.
36. Hornik, R. C., & Niederdeppe, J. (2008). Information scanning. In W. Donsbach (Ed.), Interna- tional encyclopedia of communication (pp. 2257–2261). Oxford, UK: Wiley-Blackwell.
37. Kunda, Z. (1990). The case for motivated reasoning. Psychological bulletin, 108(3), 480.
38. Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. Nature neuroscience, 23(1), 130-137.
39. Kappes, A., & Sharot, T. (2019). The automatic nature of motivated belief updating. Behavioural Public Policy, 3(1), 87-103.

40. Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition.

41. Klayman, J. & Ha, Y.-W. Confirmation, disconfirmation, and information in hypothesis testing. Psychol. Rev. 94, 211 (1987).

42. Lebreton, M., Abitbol, R., Daunizeau, J., & Pessiglione, M. (2015). Automatic integration of confidence in the brain valuation signal. Nature Neuroscience, 18(8), 1159-1167.

43. Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., ... & Luigjes, J. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. Science Advances, 4(5), eaaq0668.

44. Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*(4), 1-9.

45. Loewenstein, G., & Molnar, A. (2018). The renaissance of belief-based utility in economics. Nature Human Behaviour, 2(3), 166-167.

46. Loewenstein, G. (2006). The pleasures and pains of information. Science, 312(5774), 704-706.

47. Lee, D. G., & Daunizeau, J. (2021). Trading mental effort for confidence in the metacognitive control of value-based decision-making. Elife, 10, e63282.

48. Maier, S. U., Beharelle, A. R., Polanía, R., Ruff, C. C., & Hare, T. A. (2020). Dissociable mechanisms govern when and how strongly reward attributes affect decisions. Nature Human Behaviour, 4(9), 949-963.

49. Metcalfe, J. (2009). Metacognitive judgments and control of study. Current Directions in Psychological Science, 18(3), 159-163

50. Mobius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2011). Managing self-confidence: Theory and experimental evidence (No. w17014). National Bureau of Economic Research.

51. Moore, D. A., & Small, D. A. (2008). When it is rational for the majority to believe that they are better than average. Rationality and social responsibility: Essays in honor of Robyn Mason Dawes, 141-174.

52. Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human frontal–subcortical circuit and asymmetric belief updating. Journal of Neuroscience, 35(42), 14077-14085.

53. Meyniel, F., Schlunegger, D., & Dehaene, S. (2015). The sense of confidence during probabilistic learning: A normative account. PLoS Comput Biol, 11(6), e1004305.

54. Meyniel, F. (2020). Brain dynamics for confidence-weighted learning. PLoS computational biology, 16(6), e1007935.

55. Noguchi, T., & Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. Psychological review, 125(4), 512.

56. Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: a randomized trial. Pediatrics, 133(4), e835-e842.

57. Oster, E., Shoulson, I., & Dorsey, E. (2013). Optimal expectations and limited medical testing: evidence from Huntington disease. American Economic Review, 103(2), 804-30.

58. Pew Research Centre, (2020).U.S. Teens Take After Their Parents Religiously, Attend Services Together and Enjoy Family Rituals.

59. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. Nature, 592(7855), 590-595.

60. Pessiglione, M., Petrovic, P., Daunizeau, J., Palminteri, S., Dolan, R. J., & Frith, C. D. (2008). Subliminal instrumental conditioning demonstrated in the human brain. Neuron, 59(4), 561-567

61. Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. Nature Neuroscience, 19(3), 366-374.

62. Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. Nature reviews neuroscience, 9(7), 545-556.

63. Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. Nature communications, 11(1), 1-11.

64. Rollwage, M., & Fleming, S. M. (2021). Confirmation bias is adaptive when coupled with efficient metacognition. Philosophical Transactions of the Royal Society B, 376(1822), 20200131.

65. Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. Proceedings of the National Academy of Sciences, 111(33), 12252-12257.

66. Schulz, L., Rollwage, M., Dolan, R. J., & Fleming, S. M. (2020). Dogmatism manifests in lowered information search under uncertainty. Proceedings of the National Academy of Sciences, 117(49), 31527-31534.

67. Schulz, L., Fleming, S. M., & Dayan, P. (2021). Metacognitive Computations for Information Search: Confidence in Control. bioRxiv.

68. Sharot, T. & Garrett, N. (2016) Forming Beliefs: Why Valence Matters. Trends in Cognitive Sciences, 20(1), 25-33.

69. Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. Nature neuroscience, 14(11), 1475-1479.

70. Sharot, T., Velasquez, C. M., & Dolan, R. J. (2010). Do decisions shape preference? Evidence from blind choice. Psychological science, 21(9), 1231-1235.

71. Sharot, T., & Sunstein, C. R. (2020). How people decide what they want to know. Nature Human Behaviour, 4(1), 14-19.

72. Stephens-Davidowitz, S., & Pabon, A. (2017). Everybody lies: Big data, new data, and what the internet can tell us about who we really are. New York: HarperCollins.

73. Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., & Sharot, T. (2016). How people update beliefs about climate change: Good news and bad news. Cornell L. Rev., 102, 1431.

74. Sunstein, C. R. (2019). Ruining popcorn? The welfare effects of information. Journal of Risk and Uncertainty, 58(2), 121-142.

75. Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracín, D. (2015). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. Psychological Bulletin, 141(6), 1178–1204.

76. Tappin, B. M., Van Der Leer, L., & McKay, R. T. (2017). The heart trumps the head: Desirability bias in political belief revision. Journal of Experimental Psychology: General, 146(8), 1143.

77. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. science, 185(4157), 1124-1131.

78. Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K., & Tucker, J. A. (2020). Political psychology in the digital (mis) information age: A model of news belief and sharing.

79. Van Bavel, J. J., Sternisko, A., Harris, E. A., & Robertson, C. (2019). The social function of rationalization: An identity perspective.