The rational dynamics of implicit thought

Brett Karlan
University of Pittsburgh

This is a pre-print of a paper to be published in the Australasian Journal of Philosophy. *Please cite the published version when possible.*

ABSTRACT

Implicit attitudes are mental states posited by psychologists to explain behaviors including implicit racial and gender bias. In this paper I investigate the belief view of the implicit attitudes, on which implicit attitudes are a kind of implicit belief. In particular, I focus on why implicit attitudes, if they are beliefs, are often resistant to updating in light of new evidence. I argue that extant versions of the belief view do not give a satisfactory account of this phenomenon. This is because proponents of the belief view have focused on overly narrow explanations of recalcitrance in terms of belief storage. Expanding the focus of the belief view to the kinds of irrational and arational transitions between beliefs and other mental states provides compelling (if preliminary) explanations of recalcitrance.

1. Introduction

Consider John:

> ***Bias.*** John seems to sincerely endorse the claim that women and men are both equally cut out for the work at his company. Nonetheless, when given a choice between similarly qualified male and female candidates, John usually hires the man. When presented with evidence of the quality of work that women can do for his company, his behavior does not change.

There is a disconnect between what John believes (or at least claims to believe) and his behavior. He seems to be a card-carrying egalitarian about gender in the workplace, but his behavior mirrors that of an old-school sexist. John might be lying about what he believes, of course.[1] But assume he is not. Another possibility is psychologically more interesting. John might believe that women and men are equally well cut out for the work, but also might harbor some other attitude that conflicts with his beliefs and drives his behavior.

There is a significant amount of work in social and cognitive psychology dedicated to this possibility, where the second, discordant attitude is referred to as an *implicit attitude*.

[1] Explicit sexism is an important causal driver of workplace inequality [Koch et al. 2014], and the following discussion is not meant to minimize the role that explicit prejudice plays in discrimination.

Psychologists posit implicit attitudes in order to give satisfactory explanations for a range of dissonance cases:

- Explicitly egalitarian instructors who show negative implicit attitudes towards Black students are more likely to grade them harshly [Jacoby-Senghor et al. 2016].

- Healthcare professionals with negative implicit biases towards minority groups are more likely to minimize their pain and provide them with substandard treatment [Chapman et al. 2013].

- Consumers with no explicit preferences for products can develop implicit attitudes towards a brand that strongly influences the decision to buy (say) toothpaste under time pressure [Friese et al. 2006].

Philosophers of mind have developed a litany of proposals for making sense of the implicit attitudes. *Inter alia*, it has been proposed that implicit attitudes are constituted by associations between concepts [Holroyd 2016; Byrd 2019], aliefs [Gendler 2008; Madva 2016], beliefs [Mandelbaum 2016; Egan 2011; Frankish 2016], affective tension clusters [Brownstein 2018], character traits [Machery 2016], the outputs of three separate evaluation systems [Huebner 2016], reasoning structures without corresponding states [Johnson 2020], imaginings [Welpinghus 2020; Sullivan-Bissett 2019], mental imagery [Nanay forthcoming], and *sui generis* cognitive states like patchy endorsements [Levy 2015] and in-between beliefs [Schwitzgebel 2010]. Understanding the character and function of implicit attitudes is important, both for the philosophical project of accounting for the states that furnish the mind, and for normative projects that attempt to evaluate agents and their actions, where the influence of implicit attitudes may loom large.[2]

[2] On the way I am using the term 'implicit attitude,' an implicit attitude is whatever mental state (or suite of mental states) explains behaviors like ***Bias***. While it often seems that these states are inaccessible or unconscious in many cases, my account does not define implicit attitudes in terms of their inaccessibility (see, in contrast, Mandelbaum [2016]). I do this, in part, because there is significant controversy concerning whether or not implicit attitudes are, in fact, consciously inaccessible [Gawronski et al. 2006; Toribio 2018].

In this paper I focus on the *belief view*. This view says that the implicit attitudes are best thought of as a species of the familiar psychological kind BELIEF. John may explicitly believe in equal treatment for men and women in the workplace, but he also harbors some other belief (an *implicit belief*) that says otherwise. In slogan form: implicit attitudes are implicit beliefs. While the belief view has been bolstered by recent results in implicit attitude research, I think a central question for the belief view has yet to receive a satisfying answer: if implicit attitudes are beliefs, why are they often radically resistant to rational updating in light of new evidence? I argue that previous attempts to reconcile evidence recalcitrance with the belief view are less than satisfactory. But I argue this is partially because proponents of the belief view have focused narrowly on explaining recalcitrance in terms of architecture for belief storage. If, instead, we focus on inferential transitions between beliefs and other mental states, I argue the belief view looks much more plausible. While these explanations await further empirical verification, they provide us with strong (initial) reason to endorse the belief view.

2. The belief view

While there are several differing versions of the belief view,[3] they can be characterized by a common core idea:

> ***Belief view.*** In central cases (especially ones similar to ***Bias***), implicit attitudes are implicit beliefs.

The belief view explains John's behavior like so: John may genuinely believe that men and women are equally cut out for the work, but he also harbors a different kind of belief, one that is not so egalitarian. This implicit belief, developed over the course of years of being exposed to a culture that tends to portray women as homemakers and men as breadwinners, influences his

[3] Prominent proponents of the belief view of the implicit attitudes include Mandelbaum [2016], Egan [2011], Frankish [2016], De Houwer [2014], and Gertler [2011].

hiring decisions. His implicit belief might crowd out his more explicit egalitarian commitments in several ways: it might be active under cognitive load or distraction [Wigboldus et al. 2004], or it might primarily function as a tiebreaker when candidates otherwise seem relatively similarly qualified [Ulhmann and Cohen 2005]. Nonetheless, implicit beliefs can influence our behavior in a host of pernicious, if subtle, ways.

This kind of explanation, in virtue of invoking a reasons-responsive propositional attitude like belief, proposes that implicit attitudes operate in *rational* ways. This stands in sharp contrast to another kind of explanation that is often given for implicit bias, on which the functioning of implicit attitudes is *arational*.[4] An (oversimplified) model: John tends to associate the concept MAN with WORK, and the concept WOMAN with HOME. Because of this association, he is much more likely to think of MAN when in WORK-priming environments, and it is this disposition that makes him more likely to hire the man for the job. On these kinds of explanations, the relations between implicit attitudes, their environment, and behavior is *merely causal*: associations between concepts do not update in response to evidence, nor do they enter into inferences with other mental states to influence behavior.[5] In short, the opponents of the belief view are committed to a dissimilarity between the functional roles of implicit attitudes and the canonical functional roles of belief.

Mandelbaum [2016] offers a litany of evidence in support of the belief view, focusing on making the positive case that implicit attitudes have the capacity to respond to evidence. Mandelbaum argues that implicit attitudes seem to engage in something similar to 'the enemy of my enemy is my friend' reasoning [Gawronski et al. 2005]: subjects who are negatively

[4] Using the term 'rational' in this way entails that the complement of a rational mental state is an arational mental state, not an irrational one. Implicit attitudes can be well-attuned or mis-attuned to the evidence, but what makes them rational (in this sense) is their capacity to respond to evidence, in contrast to those states that cannot respond at all.

[5] Accounts of implicit bias that entail their arationality include Gendler [2008; 2011], Madva [2016], Brownstein [2018], and Holroyd [2016].

implicitly biased against some group, and learn that others are similarly negatively biased, tend to develop a positive implicit attitude towards this second group. This is easy to explain if implicit attitudes can track evidence-based conditions like who is disposed to like whom, but hard to explain if implicit attitudes are mere concatenations of arational associations. He also argues that implicit attitudes are directly modulated by the strength of arguments [Briñol et al. 2009], and can be formed on the basis of abstract learning rather than repeated conditioning [Gregg et al. 2006]. These pieces of evidence point towards a rational, rather than merely causal, pattern of function for implicit attitudes.

More recent research offers further evidence that implicit attitudes can update in response to evidence in paradigmatically belief-like ways. Implicit impressions of the trustworthiness of others' faces, for instance, can be updated when countervailing evidence is presented to the subject, provided the evidence is marked and diagnostic [Shen, Mann, & Ferguson 2020]. Implicit attitudes about the behavior of moral agents track the same fixation and modulation mechanisms as explicit beliefs about the subject, a mixture of observable behavior and inferred mental states of agents [Kurdi, Krosch, and Ferguson 2020]. In particular, this body of work has shown that properties of presented information that make a difference to the rational character of that information also make a difference to whether subjects update their implicit attitudes. Whether a piece of information is *diagnostic* of the content of the implicit attitude (that is, revealing of the nature of that thing) and whether it is *believable* (roughly, whether it coheres with how the subject otherwise approaches the content) are both predictive of whether subjects immediately change their attitudes when presented with new evidence [Ferguson et al. 2019] and whether those changes persist [Cone, Flaharty, and Ferguson 2021]. In other words, implicit attitudes seem to have the capacity to (and very often do!) update in light of new evidence in much the same way as beliefs.

This empirical story has not gone unchallenged. Some of the studies and interpretations that Mandelbaum presents in his paper have been questioned. Briñol et al. [2009], for instance, has been criticized for being statistically underpowered in a way that undermines the possible import the study might have for the belief view (see the methodological discussions in Byrd [2019]). Some have also thought the interpretations Mandelbaum and others have given of empirical results tend to be objectionably selective. Mandelbaum's interpretation of Gawronski et al. [2005], for example, as supporting a kind of 'enemy-of-my-enemy' reasoning, is in direct contradiction with the interpretation that the authors themselves give of their own work. And there remain many in the empirical literature skeptical of the ability of evidence to change our implicit attitudes (Lai et al. [2016]), though more recent work presents a picture more favorable to the belief theorist [Mann and Ferguson 2015; Cone, Flaharty, and Ferguson 2021]. For our purposes, it will be sufficient to note that the proper interpretation of the empirical literature surrounding the belief view is one that remains significantly controversial, and neither the belief view nor associationism should be considered the default interpretation of this literature.[6]

All of these issues are ones that, ultimately, a belief theorist must deal with, though there are reasons to be hopeful that such a response is in the offing (see Bendana & Mandelbaum [forthcoming]). In any case, this is not the line of argument against the belief view I want to focus on in this paper.[7] I instead want to focus on a different kind of response to the belief view, one that has been undertheorized until now. This response admits that implicit attitudes sometimes function in rational ways, but points to other cases where implicit attitudes seem to function in strikingly irrational ways. While implicit attitudes might be propositional attitudes

[6] My thanks to an anonymous reviewer for pushing me on some of these points.
[7] Indeed, I believe that even if all of the concerns raised above are valid ones, the remainder of the paper is still of significant interest. The project would then be a conditional one, asking the reader to temporarily accept the belief view (whatever its flaws) and see whether it can deal with a heretofore underexplored problem. I think it can, and think this success actually gives us resources to make sense of some of the controversy already discussed. I am thankful to an anonymous reviewer for pushing me to make this way of approaching the paper more clear.

of some kind, why specifically think that they are *beliefs*, given their shocking resistance to updating in many cases [Levy 2015]? If one has to go so far as undertaking weeks of love-and-kindness meditation to undo the effects of an implicit attitude [Kang et al. 2014], why think that implicit attitude is anything like what we normally think of as BELIEF? These theorists develop alternative propositional accounts of implicit attitudes that seek to explain the same data as the belief view, but with the added benefit of being able to make sense of evidence insensitivity in a way the belief view cannot. I think this objection is a strong one, and the proponent of the belief view should have something to say in response.

3. The problem of recalcitrance

The belief view of the implicit attitudes seems implausible to many, in part, because of significant evidence of *recalcitrance* in implicit attitudes. Beliefs are mental states that are paradigmatically evidence sensitive: in general, beliefs tend to update in light of new evidence.[8] Even if the evidence marshalled in Section 2 demonstrates that implicit attitudes often do respond to evidence in belief-like ways, beliefs do *not* tend to paradigmatically function in some of the odd ways that implicit attitudes often function, and this cries out for explanation. Some representative findings:

- Subjects who demonstrate implicit biases against Black candidates for a professorship, when given arguments for the effectiveness and moral worth of affirmative action, not only tend to ignore that evidence, but often end up demonstrating more implicitly biased behaviors against Black candidates as a result [Gawronski et al. 2008].

- Subjects with anti-Black implicit attitudes often have to go through extensive therapies, such as love-and-kindness meditation [Kang et al. 2014], in order to stably change their implicit attitudes.

- Merely being exposed to the words 'Muslim' and 'terrorist,' even when reviewing statements like 'it is wrong to think of all Muslims as terrorists,' is sufficient to cause an increase in anti-Muslim implicit attitudes in subjects [Deutsch and Strack 2010].

[8] The exact modal operator needed to make sense of this tendency is controversial; see Shah and Velleman [2005] and Helton [2020] for some differing views.

It is incumbent on proponents of the belief view to be able to account for this recalcitrance. If implicit attitudes are beliefs, why do they not function the way we expect beliefs to function in these, and many other, cases (cf. Madva [2016])?

One response denies that beliefs, in general, have to have the capacity to respond to evidence in order to truly be beliefs [Viedge 2018]. Just as some argue that delusions might be beliefs without having the capacity to ever respond to evidence, so too might implicit attitudes.[9] Alternatively, perhaps for a mental state to be a belief, it is necessary for it to belong to a mental *kind* that is paradigmatically evidence sensitive, even if individual *instances* of the kind might completely lack this sensitivity. While there is an interesting metaphysical debate to be had here about necessary conditions for belief, I think this is an unsatisfying way to account for the nature of the implicit attitudes for at least two reasons. First, this view of belief is non-standard in the philosophy of mind. A defense of the belief view of the implicit attitudes that begins by adopting a view of belief that many would not accept loses much of the motivation for assimilating implicit attitudes to a more familiar psychological category in the first place. Additionally, given my theoretical interest in the rational dynamics of implicit thought, this move is particularly unsatisfying, since it removes some beliefs from the realm of rational evaluation. If a mental state lacks even the capacity to respond to evidence, how can it be rational or irrational for subjects to hold that belief (cf. Helton [2020])? For these reasons, I will place the evidence-resistant view of belief to the side going forward.

A more popular approach for explaining the evidence-insensitivity of implicit beliefs centers on the idea of *mental fragmentation*. The basic idea of fragmentation involves rejecting the

[9] The literature on delusions, in fact, represents a great comparison case for our purposes, since doxasticists about delusion also must make sense of significant evidence of recalcitrance in delusions. Flores [2021] presents a compelling argument, similar in both content and spirit to the current argument, in the delusion context: the purported evidence recalcitrance of delusions is not evidence against doxasticism, since there are many plausible masking mechanisms that explain why a particular delusion might not update in a particular case. This paper adopts a similar approach for implicit attitudes.

idea that there is a single interconnected web of belief that subjects maintain. Instead, subjects have several different webs of belief, desires, and intentions (collectively called *fragments*), and there are importantly different coherence properties between and within fragments. Beliefs that lay within the same fragment will tend to be coherent with one another, but there is no coherence requirement for beliefs across fragments. Modeling beliefs in this way allows us to make sense, for instance, of Lewis [1982]'s self-report that he seemed to believe (i) that Nassau Street runs east-west, (ii) that the train tracks run north-south, and (iii) that the train tracks and Nassau Street run in parallel. Each belief occupies its own belief fragment, and since they are very rarely active at the same time, there is no pressure for interfragment coherence. A similar story can be told about explicit and implicit beliefs: they each occupy a different belief fragment, and there is often little coherence between them. John may explicitly believe that men and women are equally cut out for the work, but this belief is powerless when his implicit fragment is active.[10]

As an account of the mental architecture of humans, the fragmentation account seems plausible, and I argue below that a version of the account can explain some of the recalcitrance of implicit attitudes. But I think existing fragmentation accounts fail to offer us a satisfying account of the recalcitrance of implicit beliefs. They offer little explanation why sometimes the *same* implicit attitudes respond to evidence well and other times do so poorly. On one version of the fragmentation view, the *interpretationist* version, this is by design [Elga and Rayo forthcoming]. Interpretationist fragmentation theory takes a subject's recalcitrant irrationality as a datum to be incorporated into a systematized theory that can satisfactorily capture a subject's behavior. These theories are explanatory in some sense: interpertationist fragmentation theories 'identify patterns and show that relevant facts are instances of those patterns' [Elga and Rayo

[10] The general theory of fragmentation is developed by, *inter alia*, Stalnaker [1984], Stitch [1990], and Greco [2014].

forthcoming: 6]. But they do not offer an account of the origins of the patterns, since they aim to be theoretically neutral about the underlying architecture that might produce behavior. Given the goals of this kind of fragmentation theory, it is not an objection to point out that the theory merely systematizes a subject's response. But this also means that interpretationist fragmentation theory cannot give us a theory of recalcitrance that will help defend the belief view from its opponents [Norby 2014].

The *realist* version of fragmentationism, defended by Bendana and Mandelbaum [forthcoming], in contrast, does attempt to give an explanation of the rational function of different fragments.[11] On their way of understanding implicit attitude formation and revision, implicit beliefs about social groups are housed in many psychologically real, redundant fragments that are tokened in particular contexts. When John considers the arguments for equality in the workplace, he creates a fragment that believes that men and women are equal. But when he implicitly encodes stereotypes about women being homemakers, he encodes this belief in a different fragment. Working hard in one context to overcome a particular attitude will not tend to change other beliefs that are stored in other fragments. This, in turn, is meant to explain why short-term attempts to change implicit attitudes often fail, while long-term strategies are often more successful.

On Bendana and Mandelbaum's formulation, fragments obey what they call the Environmental Principle, which states that:

> novel fragments are opened up in novel environments. According to this principle, when one is visiting Spain for the first time, one opens up a new fragment with SPAIN as the heading. Of course, one doesn't just visit Spain, one goes to the Madrid airport or the Sagrada Familia. For each of these places, we assume that a new fragment will be opened [Bendana and Mandelbaum forthcoming: 29].

[11] For instance: 'fragmentation can... explain seemingly disparate social psychological findings regarding implicit bias' [Bendana and Mandelbaum forthcoming: 10].

It follows from the hyperspecificity of the Environmental Principle that, in general, there will be no story to tell about why some beliefs are held in response to evidence and others are not.[12] That is just the way those beliefs were encoded in their particular contexts. The concept of fragmentation, at least in its Environmental Principle formulation, cannot explain either evidence sensitivity or recalcitrance of a single implicit attitude. The fact that one and the same implicit attitude seems to be evidence sensitive in some contexts and radically resistant to evidential updating is only explained, if it is explained at all, in terms of two *different* attitudes being tokened at two different times, with one of the tokenings being responsive but not another. There is no further fact to be had about the rational response of implicit attitudes. But this is just to leave the crucial question of recalcitrance unanswered.

4. Belief, inference, transition

The two most thorough extant accounts of fragmentation, then, fail to bolster the belief view, either by methodological design or because of hyperspecificity of fragment individuation. These approaches attempt to explain implicit attitude recalcitrance in terms of belief storage mechanisms: implicit beliefs do not respond to evidence or other rational pressure because they are housed in fragments without strong connections to other fragments. Focusing only on storage, however, threatens to unnecessarily hamper the explanatory resources available to the belief view. In this section, I argue that focusing on mental transitions, especially rationally evaluable inferential transitions, gives the belief view a new host of explanatory possibilities for making sense of evidence recalcitrance in implicit attitudes.[13]

[12] This is an instance of a general worry for fragmentation theory that Elga and Rayo raise, arguing that the theory must ensure that 'access table's elicitation conditions are not individuated too finely. Otherwise, an access table might become a mere listing of overly specific dispositions, and so fail to provide useful explanations of behavior' [Elga and Rayo forthcoming: 6].

[13] Despite my critique of the extreme fragmentation account he prefers, my approach represents a clear continuation of the Mandelbaum [2016] project of laying out the belief view. Mandelbaum discusses evidence of genuine evidential updating, arguing the arationality approach is mistaken and that implicit

4.1 The messiness of belief

Pre-theoretically, one might think belief states are those that *only* enter into rational transitions representable by simple logical deductions. But this idealized picture is hopeless for making sense of human belief systems, where rational transitions between premises and conclusions sit next to all manner of irrational and arational belief-involving transitions. As an example, one such effect is the tendency for beliefs to enter into associative relations in virtue of their contents [Johnston 1995]. If I believe that Arsenal will win the Cup next year, I will tend to associate ARSENAL with CUP in a way I would not with another concept like CHAIR. And this will in turn make it more likely that I show classic associative behaviors, being quicker to activate one of the concepts when the other is activated.

Another transition, particularly important for the argument I make here, comes from the literature on motivated reasoning. Motivated reasoning is a broad category of mental transitions where favored conclusions are reached via methods that seek to ensure their outcomes whatever incoming evidence might be presented to the subject [Kunda 1990]. A favorite philosophical example of motivated reasoning is *wishful thinking*: a subject desires that p be true, and she thereby gathers and evaluates evidence in a way that seeks to confirm p [Siegel 2017]. Wishful thinking offers a kind of rational (as opposed to merely causal) explanation for the recalcitrance of our favored beliefs from updating: updating is blocked by certain upstream factors that influence the flow and evaluation of evidence, removing any rational pressure on the belief itself to update (by avoiding sources one thinks might provide one with counterevidence, or by developing alternative hypotheses that can explain counterevidence away).

attitudes can respond rationally to evidence. I argue that there are many arational and irrational effects on beliefs that, in turn, explain the remaining implicit attitude recalcitrance to updating.

A particular form of motivated reasoning is importantly tied to our self-conceptions and self-esteem [Mandelbaum 2019]. Most non-depressed subjects are fundamentally convinced that they are good people. This creates a strong incentive for us to find a way to discount evidence that seems to cast doubt on our fundamental goodness. Indeed, there is a litany of psychological evidence that people will go to great lengths to avoid confronting evidence of their moral transgressions [Shikta et al. 2005]. One popular strategy involves delineating which actions stem from a 'true self' from actions that can be attributed to external factors causing the agent to behave in a certain way [Newman, De Freitas, and Knobe 2015]. If I have done something reprehensible, there is significantly less rational burden on me to think of myself as a bad person if the reason I performed the action is external (for instance, if I was distracted and allowed bad advice to guide me), rather than something about my deep character. In this way, I can diffuse the rational pressure of the evidence in front of me, and my positive core belief about myself can remain unchanged.

The explanatory resources of the belief view are thus more expansive than a narrow concentration on fragmented belief storage might suggest. Even at the level of explicit belief, there are a whole host of ways that beliefs can become enmeshed in networks that produce irrational or arational transitions between belief states. I next argue the belief theorist should apply the same explanatory resources to the case of implicit belief.

4.2 Implicit wishful thinking

Building on the previous section, I here want to argue that motivated reasoning and wishful thinking can occur in the implicit as well as the explicit domain. Many of the results that push theorists away from the belief view can be explained by an interaction between affect, cognition, and motivation at the implicit level, just as such deviations can be explained in the

more explicit cases discussed above. As far as I am aware, this proposal has never been theorized or worked out in detail, so what I present here is a mere sketch. Nonetheless, I think the sketch is a compelling way to extend the belief view.

Here is a toy example to help set up the model. It is plausible (see below) that subjects not only have implicit cognitive attitudes towards others in social groups, but also implicit prejudices and animus towards them, as well as implicit self-conceptions which are usually positive. Suppose a particular white subject holds (a) an implicit belief that Black men are lazy, (b) an implicit racial animus towards Black men, and (c) a positive implicit self-conception. As it stands, (a) and (b) seem to put some rational pressure on (c): if he really is such a good person, why does he house this animus towards members of a particular group? The subject can resolve this pressure, however, if he thinks there is good evidence in favor of (a), and that this evidence justifies his prejudice in (b). This is, in fact, the result we see in racial bias research: there tends to be a mutually reinforcing network between believing that white privilege does not exist, that members of other racial groups do not get ahead because they do not work hard enough, and corresponding racial animus and positive self-esteem [Wilkins and Kaiser 2014].

Social psychologists have established the plausibility of distinguishing between implicit cognition, affect, and motivation. Different implicit attitude measures, for one thing, are thought to pick out cognitive (the Implicit Association Task [Greenwald, McGhee, and Schwartz 1998]), affective (the Implicit Positive and Negative Affect Test [Quirin, Khazen, and Kuhl 2009]), and self-esteem (a modified form of the IAT [Yamaguchi et al. 2007]) aspects of implicit thinking. Psychologists have also found evidence of interactions between these states that justify giving them the structure corresponding to their labels: for instance, when a person's self-esteem is threatened (by losing a contest in something the subject cares about, for instance), there tends to be an overcompensation to ward off possible negative self-feeling, as well as a corresponding

increase in in-group bias as a further way to dismiss the threat [Rudman, Dohn, and Fairchild 2007]. So not only does it seem we are able to pick out implicit states that have similarities to the kinds of functions we would normally pick out with the terms 'cognitive,' 'affective,' and 'motivational,' it also seems these states interact in ways similar to the ways these states interact explicitly. It is this kind of interaction that the belief view should look to exploit.

Thinking about the interaction between implicit belief, affect, and motivation can help explain many of the puzzling results we have already gestured at in this paper. The model, for instance, can explain the otherwise odd *attitudinal backfiring* of implicit attitudes cited above [Gawronski et al. 2008]. In order to discount evidence that goes against one's favored beliefs, one might explore nearby hypothesis spaces to come up with an alternative account of the evidence. Perhaps one might settle on the hypothesis that God-hating liberals are producing the so-called 'evidence' for gender equality because they hate the natural order of things. This conspiracy is very much able to explain the evidence presented of women being cut out for the work. The truth of this undermining hypothesis eliminates the rival hypothesis (that women really are cut out for the work) from consideration, and because one is now distributing one's credences over a smaller probability space, one might increase one's credence in the original hypothesis.[14]

Implicit motivated reasoning can also explain a crucial finding of Mann and Ferguson [2015]: implicit attitudes that are newly formed are much more likely to be revised in light of new evidence than those that are entrenched. Newly formed implicit attitudes are less likely to be linked with a large network of deeply entrenched networks of affect and motivation, so when new evidence is subsequently presented that the subject's view is mistaken, there is no self-regarding pressure on the subject to find a way to maintain the attitude. Once implicit

[14] For a description of this process in much more detail, see Kelly [2008].

attitudes have been incorporated and entrenched in these complicated and mutually-reinforcing networks, however, changing them becomes much harder. In a similar vein, the conflict between evidence and self-conception in implicit thought can explain why children tend to develop implicit attitudes that are (initially) more evidence sensitive than those in adults [Charlesworth et al. 2020]. Children are still developing complicated networks of self-regard during the same time they are developing implicit attitudes [Robins and Trzesniewski 2005], and the lack of an established connection between the two reduces the rational pressure that such a network can exert on a particular implicit belief that a child may develop.

It is not my goal here to give extensive empirical support to any particular explanation of implicit attitude recalcitrance in terms of motivated reasoning, though I do think these explanations are plausible. Instead, I want to point out the *kinds* of explanatory resources focusing on inferential transitions gives the proponent of the belief view. Beliefs function in any number of irrational or arational ways, and the belief theorist can use these facts to explain otherwise baffling instances of recalcitrance in implicit attitudes. The opponent of the belief view thinks there is a deep divide between the functional roles of beliefs and implicit attitudes; but when we focus on how beliefs actually function, even at the explicit level, this claim loses plausibility.

4.3 Moderate fragmentation

For most of this paper, I have been focusing on the recalcitrance of a single implicit attitude when the subject is presented with countervailing information. I have been arguing that looking at the possible inferential transitions such states can be embedded in helps to make sense of their recalcitrance. But there is a second kind of recalcitrance often present in implicit bias cases: an incoherence between explicit and implicit attitudes. It is common for subjects who

endorse egalitarian explicit commitments to nonetheless harbor implicit attitudes that contradict those commitments; this is the kind of case that motivated our reflection on ***Bias*** in the first place.[15] How should the proponent of the belief view make sense of this?

It is not immediately obvious how a focus on implicit motivated reasoning could explain this kind of case. If there is rational pressure being generated by a subject's positive self-conception, shouldn't that pressure make it more likely that they will change their implicit attitudes, especially when those attitudes are made explicit to them? What about members of minority groups who develop negative implicit biases towards their own groups [Rudman and Goodwin 2004], which seem to actively cut against their own positive self-conception? While I do not think the notion of implicit motivated reasoning can generate compelling explanations in all of these cases, some explanations are possible. To take just one example, it is likely that implicit beliefs are usually formed ballistically [Mandelbaum 2015]: subjects take in ambient information about what dominant social groups around them think of others and immediately encode this as a belief. Once that belief is present, it itself generates some rational pressure to modify other aspects of the belief-affect-self-conception network. This, in turn, might produce something more akin to a self-deception case, where subjects are in some sense aware of their implicit attitudes but tend to keep them out of deliberation as much as possible because of the extreme cognitive dissonance they produce. It is plausible that many people are in some sense aware of what they implicitly believe [Toribio 2018], so this result is not unwarranted.

Nonetheless, the belief theorist should freely admit that these explanations are not airtight or fully satisfying. Instead, I think she should instead endorse a moderate realist fragmentationism. The realist account of fragmentation given by Bendana and Mandelbaum fails, in my estimation, not because of the underlying facts about fragmentation, but because a

[15] See also the idea that implicit attitudes among dominant group members represent the 'hidden biases' of otherwise good people [Banaji and Greenwald 2013].

commitment to hyperspecificity of fragment creation undermines our ability to identify one and the same attitude across different environments. On their account, there just is no answer to the question of what a subject believes across different contexts. But one need not go in for this extreme view to see the appeal of fragmentation as a theory of human mental architecture. Indeed, it is hard to see how many of the results on human logical inconsistency [Cooper and Duncan 1971] and so-called Spinozistic believing [Gilbert 1991] can be incorporated into a single belief-web framework. So the idea that the human mind is divided up into several independent belief stores itself becomes a feature the belief theorist can use to explain evidence recalcitrance in implicit attitudes.

The explanatory resources of the belief view are thus multifaceted and intricate. In terms of belief formation and storage, the belief theorist can appeal to a certain (probably small) number of distinct fragments where beliefs can be stored, with strong intrafragment coherence but little interfragment coherence. If evidence is presented to the subject, one reason a particular fragment might not update is because the information was not made available to that fragment. This will not be plausible in many cases, however, especially in social cognition where most information will be encoded exclusively in implicit fragments. Instead, the belief theorist should appeal to a host of irrational and arational mental transitions that include belief states and that mask the rational updating function of belief *within* a particular fragment. When these masking features are removed, or when the newly produced attitude is not given sufficient time to embed in the network, we see the rational updating of these states. But recalcitrance is no surprise when dealing with long-term, entrenched networks of implicit belief, affect, and motivation.

Far from being singularly focused on belief storage, the explanatory resources of the belief view are vast, and are supported by the available empirical evidence.[16] The belief view, at the very least, represents an active and interesting research program that begets future work.

5. Conclusion

In this paper, I have tried to expand the explanatory possibilities for the belief view of the implicit attitudes. I take it there are two explanatory projects one needs to undertake if one thinks implicit attitudes are beliefs: showing that their basic function is belief-like, and having plausible explanations for purported differences between the two. Other authors have focused on providing the former; here, I lay a pathway for the latter, focusing on the transitions that occur within a particular implicit fragment. I think there are significant conceptual resources here for making sense of the evidence recalcitrance of implicit attitudes, especially since ordinary explicit beliefs are often recalcitrant in exactly the same ways. Implicit attitudes, like belief more generally, are messy.

My goal has merely been to identify and explore some of the explanatory resources available to the belief view. It remains for future work to confirm the particularities of any belief account, and the epistemic value of the account rests partially on these empirical confirmations. Some results (for instance, Charlesworth et al. [2020]) already seem to lend support, but clearly much more work needs to be done. But much like the other proposals that the belief view interacts with, I take it that the epistemic standing of the belief view rests both on empirical

[16] There is another sense in which the current approach differs from others. Often some of the results catalogued here are used to motivate the *psychological immune system* hypothesis [Mandelbaum 2019], which states that the primary function of belief is not to update in light of evidence but to produce a relatively functional person in the face of evidence of terribleness in the world and in ourselves. I do not see why the defender of the belief view needs to commit themselves to a new constitutive aim of belief, however. One can just as easily think belief does (functionally) aim at truth, and that unimpeded beliefs will update in light of evidence. In the human mind, however, few things are unimpeded. The belief view is thus separate from the psychological immune system hypothesis.

results and on the kind of philosophical and explanatory light such a view can throw on issues we care about as philosophers of cognitive science and mind. Whether we can ultimately offer an explanation that is both philosophically compelling and empirical supported will be the task of future work, but the prognosis looks promising.

Acknowledgments

I must particularly thank Thomas Kelly, Grace Helton, and Gideon Rosen for their extensive engagement with this material, which formed the core of the dissertation they examined for my PhD at Princeton University. I am also deeply grateful for discussions with Colin Allen, Colin Bradley, Haley Brennan, Sam Clarke, Adam Elga, Eleanor Gordon-Smith, Daniel Kranzelbinder, Thomas Lambert, Kyle Landrum, Jared Liebergen, Eric Mandelbaum, Camilo Martinez, Alex Meehan, Patrick Miller, Joe Moore, Alejandro Naranjo Sandoval, Chris Register, Gabriel Shapiro, Erik Zhang, and participants in several iterations of the Princeton Dissertation Seminar and Princeton Philosophical Society, all of whom helped improve the paper considerably. Thanks is also due to audiences at the Oxford Graduate Conference (particularly my commenter Rachel Elizabeth Fraser), the Society for Philosophy and Psychology 2019 annual meeting, and the Southern Society for Philosophy and Psychology 2019 annual meeting (particularly my commenter Jessica Wright). Finally, thanks to two anonymous referees for thorough and helpful revisions.

References

Banaji, M.R. and A.G. Greenwald 2013. *Blindspot: Hidden Biases of Good People*. New York: Delacorte Press.

Bendana, J. and E. Mandelbaum forthcoming. The Fragmentation of Belief, in *The Fragmented Mind*, ed. Cristina Borgoni, Dirk Kindermann, and Andrea Onofri, Oxford: Oxford University Press.

Briñol, P., R. Petty, and M. McCaslin 2009. Changing Attitudes on Implicit versus Explicit

Measures: What is the Difference? in *Attitudes: Insights from the New Implicit Measures*, ed. Russell H. Fazio, Pablo Briñol, Richard E. Petty, New York: Psychology Press: 285-326.
Brownstein, M. 2018. *The Implicit Mind: Cognitive Architecture, the Self, and Ethics*. Oxford: Oxford University Press.
Byrd, N. 2019. What We Can (And Can't) Infer About Implicit Bias From Debiasing Experiments. *Synthese* 198/1: 1-29.
Chapman, E.N., A. Kaatz, and M. Carnes 2013. Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities. *Journal of General Internal Medicine* 28/11: 1504-10.
Charlesworth, T.E., B. Kurdi, and M.R. Banaji 2020. Children's Implicit Attitude Acquisition: Evaluative Statements Succeed, Repeated Pairings Fail. *Developmental Science* 23/3: 1-10.
Cone, J., K. Flaharty, and M.J. Ferguson 2021. The Long-Term Effects of New Evidence on Implicit Impressions of Other People. *Psychological Science* 33/2: 173-88.
Cooper, J., and B.L. Duncan 1971. Cognitive Dissonance as a Function of Self-Esteem and Logical Inconsistency. *Journal of Personality* 39/2: 289-302.
De Houwer, J. 2014. A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass* 8/7: 342-53.
Deutsch, R. and F. Strack 2010. Building Blocks of Social Behavior: Reflective and Impulsive Processes, in *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*, ed. Bertram Gawronski, and B. Keith Payne, New York: The Guilford Press: 62–79.
Egan, A. 2011. Comments on Gendler's, "The Epistemic Costs of Implicit Bias." *Philosophical Studies* 156/1: 65.
Elga, A. and A. Rayo forthcoming. Fragmentation and Information Access, in *The Fragmented Mind*, ed. Cristina Borgoni, Dirk Kindermann, and Andrea Onofri, Oxford: Oxford University Press.
Ferguson, M.J., T.C. Mann, J. Cone, and X. Shen 2019. When and How Implicit First Impressions Can Be Updated. *Current Directions in Psychological Science* 28/4: 331-36.
Flores, C. 2021. Delusional Evidence-Responsiveness. *Synthese*: 1-32.
Frankish, K. 2016. Playing Double: Implicit Bias, Dual Levels, and Self-Control, in *Implicit Bias and Philosophy*, ed. Michael Brownstein and Jennifer Saul, Oxford: Oxford University Press: 23-46.
Friese, M., M. Wänke, and H. Plessner 2006. Implicit Consumer Preferences and their Influence on Product Choice. *Psychology & Marketing* 23/9: 727-40.
Gawronski, B., E. Walther, and H. Blank 2005. Cognitive Consistency and the Formation of Interpersonal Attitudes: Cognitive Balance Affects the Encoding of Social Information. *Journal of Experimental Social Psychology* 41/6: 618-26.
Gawronski, B., W. Hofmann, and C.J. Wilbur 2006. Are "Implicit" Attitudes Unconscious? *Consciousness and Cognition* 15/3: 485-99.
Gawronski, B., R. Deutsch, S. Mbirkou, B. Seibt, and F. Strack 2008. When "Just Say No" is Not Enough: Affirmation versus Negation Training and the Reduction of Automatic Stereotype Activation. *Journal of Experimental Social Psychology* 44/1: 370–77.
Gendler, T. S. 2008. Alief and Belief. *Journal of Philosophy* 105/10:634-63.
Gendler, T. S. 2011. On the Epistemic Cost of Implicit Bias. *Philosophical Studies* 156/1: 33-63.
Greco, D. 2014. Iteration and Fragmentation. *Philosophy and Phenomenological Research* 88/1: 656-73.
Gertler, B. 2011. Self-Knowledge and the Transparency of Belief, in *Self-Knowledge*, ed. Anthony Hatzimoysis, Oxford: Oxford University Press: 125–45.
Gilbert, D.T. 1991. How Mental Systems Believe. *American Psychologist* 46/2: 107-19.

Greenwald, A.G., D.E. McGhee, and J.L Schwartz 1998. Measuring Individual Differences in Implicit Cognition: the Implicit Association Test. *Journal of Personality and Social Psychology* 74/6: 1464-80.

Gregg, A., B. Seibt, and M. Banaji 2006. Easier Done than Undone: Asymmetry in The Malleability of Implicit Preferences. *Journal of Personality and Social Psychology* 90/1: 1–20.

Helton, G. 2020. If You Can't Change What You Believe, You Don't Believe It. *Noûs* 54/3: 501-26.

Holroyd, J. 2016. VIII- What Do We Want from a Model of Implicit Cognition? *Proceedings of the Aristotelian Society* 116/2: 153-79.

Huebner, B. 2016. Implicit Bias, Reinforcement Learning, and Scaffolded Moral Cognition, in *Implicit Bias and Philosophy*, ed. Michael Brownstein and Jennifer Saul, Oxford: Oxford University Press: 47-79.

Jacoby-Senghor, D.S., S. Sinclair, and J.N. Shelton 2016. A Lesson in Bias: The Relationship Between Implicit Racial Bias and Performance in Pedagogical Contexts. *Journal of Experimental Social Psychology* 63/1: 50-5.

Johnson, G.M. 2020. The Structure of Bias. *Mind* 129/516: 1193-236.

Johnston, M. 1995. Self-Deception and the Nature of Mind, in *Philosophy of Psychology: Debates on Psychological Explanation*, ed. Cynthia MacDonald and Graham MacDonald, Cambridge: Blackwell: 63-91.

Kang, Y., J.R. Gray, and J.F. Dovidio 2014. The Nondiscriminating Heart: Loving-Kindness Meditation Training Decreases Implicit Intergroup Bias. *Journal of Experimental Psychology: General* 143/3: 1306-13.

Kelly, T. 2008. Disagreement, Dogmatism, and Belief Polarization. *Journal of Philosophy* 105/10: 611-33.

Koch, A.J., S.D. D'Mello, and P.R. Sackett 2015. A Meta-Analysis of Gender Stereotypes and Bias in Experimental Simulations of Employment Decision Making. *Journal of Applied Psychology* 100/1: 128-61.

Kunda, Z. 1990. The Case for Motivated Reasoning. *Psychological Bulletin* 108/3: 480-498.

Kurdi, B., A.R. Krosch, and M.J. Ferguson 2020. Implicit Evaluations of Moral Agents Reflect Intent and Outcome. *Journal of Experimental Social Psychology* 90/1: 1-12.

Lai, C.K., A.L. Skinner, E. Cooley, S. Murrar, M. Brauer, T. Devos, J. Calanchini, Y.J. Xiao, C. Pedram, CK. Marshburn, and S. Simon 2016. Reducing Implicit Racial Preferences: II. Intervention Effectiveness Across Time. *Journal of Experimental Psychology: General* 145/8: 1001-16.

Levy, N. 2015. Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Noûs* 49/4: 800-23.

Lewis, D.K. 1982. Logic for Equivocators. *Noûs* 16/3: 431-41.

Machery, E. 2016. De-Freuding Implicit Attitudes, in *Implicit Bias and Philosophy*, ed. Michael Brownstein and Jennifer Saul, Oxford: Oxford University Press: 104-29.

Madva, A. 2016. Why Implicit Attitudes are (Probably) not Beliefs. *Synthese* 193/8: 2659–84.

Mandelbaum, E. 2015. The Automatic and the Ballistic: Modularity Beyond Perceptual Processes. *Philosophical Psychology* 28/8: 1147-56.

Mandelbaum, E. 2016. Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs* 50/3: 629-58.

Mandelbaum, E. 2019. Troubles with Bayesianism: An Introduction to the Psychological Immune System. *Mind & Language* 34/2: 141-57.

Mann, T.C. and M.J. Ferguson 2015. Can We Undo our First Impressions? The Role of Reinterpretation in Reversing Implicit Evaluations. *Journal of Personality and Social Psychology* 108/6: 823-49.

Nanay, B. forthcoming. Implicit Bias as Mental Imagery. *Journal of the American Philosophical Association.*

Norby, A. 2014. Against Fragmentation. *Thought: A Journal of Philosophy* 3/1: 30-8.

Newman, G.E., J. De Freitas, and J. Knobe 2015. Beliefs About the True Self Explain Asymmetries Based on Moral Judgment. *Cognitive Science* 39/1: 96-125.

Quirin, M., M. Kazén, and J. Kuhl 2009. When Nonsense Sounds Happy or Helpless: the Implicit Positive and Negative Affect Test (IPANAT). *Journal of Personality and Social Psychology* 97/3: 500-16.

Robins, R.W. and K.H Trzesniewski 2005. Self-Esteem Development Across the Lifespan. *Current Directions in Psychological Science* 14/3: 158-62.

Rudman, L.A., M.C. Dohn, and K. Fairchild 2007. Implicit Self-Esteem Compensation: Automatic Threat Defense. *Journal of Personality and Social Psychology* 93/5: 798-813.

Rudman, L.A. and S.A. Goodwin 2004. Gender Differences in Automatic In-Group Bias: Why Do Women Like Women More than Men Like Men? *Journal of Personality and Social Psychology* 87/4: 494-509.

Schwitzgebel, E. 2010. Acting Contrary to our Professed Beliefs; or the Gulf between Occurrent Judgment and Dispositional Belief. *Pacific Philosophical Quarterly* 91/4: 531-53.

Shah, N. and J.D Velleman 2005. Doxastic Deliberation. *Philosophical Review* 114/4: 497-534.

Shen, X., T.C. Mann, and M.J. Ferguson 2020. Beware a Dishonest Face?: Updating Face-Based Implicit Impressions using Diagnostic Behavioral Information. *Journal of Experimental Social Psychology* 86/1: 1-19.

Skitka, L.J., C.W. Bauman, and E.G. Sargis 2005. Moral Conviction: Another Contributor to Attitude Strength or Something More? *Journal of Personality and Social Psychology* 88/6: 895-917.

Siegel, S. 2017. How Is Wishful Seeing Like Wishful Thinking? *Philosophy and Phenomenological Research* 95/2: 408-35.

Stalnaker, R. (1984). *Inquiry*. Cambridge: Cambridge University Press.

Stich, S. 1990. *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation*. Cambridge, Massachusetts: MIT Press.

Sullivan-Bissett, E. 2019. Biased by Our Imaginings. *Mind and Language* 34/5: 627-647.

Toribio, J. 2018. Accessibility, Implicit Bias, and Epistemic Justification. *Synthese* 198/1: 1529–47.

Uhlmann, E.L. and G.L. Cohen 2005. Constructed Criteria: Redefining Merit to Justify Discrimination. *Psychological Science* 16/6: 474–480.

Viedge, N. 2018. Defending Evidence-Resistant Beliefs. *Pacific Philosophical Quarterly* 99/3: 517-37.

Wilkins, C. and C.R. Kaiser, 2014. Racial Progress as Threat to the Status Hierarchy: Implications for Perceptions of Anti-White Bias. *Psychological Science* 25/2: 439–46.

Welpinghus, A. 2020. The Imagination Model of Implicit Bias. *Philosophical Studies* 177/1: 1611–33.

Wigboldus, D.H., J.W. Sherman, H.L Franzese, and A.V. Knippenberg 2004. Capacity and Comprehension: Spontaneous Stereotyping under Cognitive Load. *Social Cognition* 22/3: 292-309.

Yamaguchi, S., A.G. Greenwald, M.R. Banaji, F. Murakami, D. Chen, K. Shiomura, C. Kobayashi, H. Cai, and A. Krendl 2007. Apparent Universality of Positive Implicit Self-Esteem. *Psychological Science* 18/6: 498-500.