Playing Double: Implicit Bias, Dual Levels, and Self-Control *

Keith Frankish

Implicit bias is sometimes thought of as a surprising discovery, uncovered by recent experimental work in social psychology. It is true that much important experimental work has been done recently, but there is a wider context to discussions about implicit bias, involving ideas about the duality of the human mind that have been around for a long time. The idea that some mental processes operate outside consciousness is — I shall argue — part of our everyday (or 'folk') conception of the mind, and implicit bias can be seen as involving a familiar phenomenon which I shall call 'playing double'. This chapter summarizes this context and draws on it to sketch a theoretical framework for thinking about implicit bias and how we can control it.

The chapter is in three parts. The first looks at implicit bias in everyday life. It begins by introducing an example of implicit bias and discussing how it contrasts with explicit bias. It then locates implicit bias within a pattern of everyday talk about implicit mentality and argues that systematic implicit bias is best thought of as manifesting a form of belief. The second part looks at the dissonance characteristic of many cases of implicit bias — cases where a person's implicit beliefs appear to conflict with their explicit ones. It asks whether such conflict is real, setting out a sceptical worry about the very existence of explicit belief, and goes on to sketch an account of explicit belief as a form of commitment. The upshot is a layered picture of the human mind, with a passive implicit level supporting an active explicit one, and this 'dual-level' view is fleshed out and compared briefly with other theories of mental duality. The third part of the chapter turns to the question of how we can overcome implicit bias. We tend to identify with our explicit mental states and processes and want them to control our behaviour. But how is such self-control possible? If we are systematically biased, how can we even form unbiased beliefs? And if we can, then how can we make them effective? The dual-level view has implications for these questions, assigning a crucial role to metacognitive mental states of certain kinds. This section discusses these issues and outlines the conditions for explicit control. The chapter concludes by identifying some predictions of the proposed account.

* This is the author's preprint of a chapter forthcoming in Michael Brownstein and Jennifer Saul (eds.), *Implicit Bias and Philosophy Volume I: Metaphysics and Epistemology*, Oxford University Press. Version 2.ii 15/05/18.

1. Implicit bias

1.1 Implicit bias in real life

In the current context, a biased person is one who is disposed to judge others according to a stereotyped conception of their social group (ethnic, gender, class, and so on), rather than by their individual talents. Such a disposition displays bias since it involves a deviation from norms of fairness.[1] A person is *implicitly* biased if their behaviour manifests a stereotyped conception of this kind, even if they do not explicitly endorse the conception and perhaps explicitly reject it. The possibility of such implicit bias is a matter of ethical concern, since it means that bias may persist in an unacknowledged, 'underground' form, even when it has been explicitly repudiated.

There is now a large body of evidence for the existence of forms of implicit bias in various experimental settings (see the Introduction to this volume for references).[2] But the broad concern, I take it, is that implicit bias may affect our behaviour and judgements across a range of everyday situations, much as explicit bias might do. Eric Schwitzgebel gives a fictional example of such implicit bias. Juliet is a white American philosophy professor. She knows there is no scientific evidence for racial differences in intelligence, and she argues with sincerity for equality of intelligence, a view which also harmonizes with her liberal outlook on other matters. Yet Juliet's unreflective behaviour and judgements of individuals display systematic racial bias:

> When she gazes out on class the first day of each term, she can't help but think that some students look brighter than others — and to her, the black students never look bright. When a black student makes an insightful comment or submits an excellent essay, she feels more surprise than she would were a white or Asian student to do so, even though her black students make insightful comments and submit excellent essays at the same rate as do the others. This bias affects her grading and the way she guides class discussion. (Schwitzgebel, 2010, p.532)

Juliet's spontaneous interactions with non-students display a similar systematic bias. Schwitzgebel notes that there need be no self-deception involved in this. Juliet might be aware that she possesses this bias, and she might even take steps to counteract it, perhaps by trying to be especially generous in her assessment of black students — though, as Schwitzgebel observes, such condescension could itself be seen as indirectly manifesting the bias. Although this is a fictional example, it is, in my experience, one that people find readily comprehensible and recognizable, and the experimental data

---

[1] I take fairness to be a norm of rationality as well as a social norm. Some writers would not use the term 'bias' for deviations from merely social norms.

[2] It is still unclear what the experimental data tell us about bias in everyday life. There are many different measures of implicit attitudes, which do not correlate well with each other and may be measuring different things (e.g., Bosson et al, 2000; Olson and Fazio, 2003). Moreover, recent meta-analyses suggest that the best-known measure of implicit bias, the Implicit Association Test (IAT) is a poor predictor of real-world ethnic and racial discrimination (Oswald et al., 2013, 2015).

are worrying precisely because they suggest that cases like it may be common in real life.

Can we be more precise about what makes Juliet's bias implicit, and about how implicit bias contrasts with the explicit sort? We might say that Juliet's bias is nonconscious, or unconscious, whereas explicit bias is conscious. This needs qualifying, however. There are different things we might mean by 'conscious', and there are senses in which Juliet's racial bias *is* conscious. First, as noted, Juliet may be aware of its existence and may consciously think *that* she is racially prejudiced, though without consciously endorsing the prejudice. It might be better to say that Juliet's bias — or, rather, the mental state underpinning it — is not introspectable: she cannot report straight off that she possesses it, as she can report her explicit views, and she becomes aware of it only through observing, or being informed of, its effects on her behaviour. (Even this may be too strong, however; there is evidence that some aspects of implicit attitudes like Juliet's are introspectable; see, e.g., Gawronski et al. 2006; Hahn et al., 2014). Second, though Juliet's bias is evident primarily in behaviour that is not consciously controlled (unreflective behaviour, as I shall call it), it may also reveal itself in behaviour that is consciously controlled (reflective behaviour).[3] It affects her conscious judgements, decisions, and feelings, and Juliet may consciously perform actions that display it — for example, consciously disciplining herself to do her grading. That is, implicitly biased actions may be consciously intended, although they are not consciously intended *to be biased*.

This suggests a better characterization of the way in which Juliet's bias is nonconscious: she does not endorse it in her conscious reasoning and decision making. Although Juliet may be conscious of behaving and judging as if there are racial differences in intelligence, she does not consciously think *that* there are racial differences in intelligence. If that thought occurs to her, she rejects it. This is compatible with her having some introspective awareness of her bias, provided she does not endorse it. By contrast, explicit bias would be bias that *is* endorsed in conscious deliberation. Thus, whereas implicit bias affects both unreflective behaviour and (some) reflective behaviour, explicit bias manifests itself only in reflective behaviour.

1.2. Implicit bias as belief

The claim that we have implicit mental states runs against a philosophical tradition, often associated with Descartes, that the mind is completely transparent to itself. However, this tradition is not the only one. There is also a long history of theoretical speculation about nonconscious processes (see Frankish and Evans, 2009), and there is a firm commonsense basis to the notion of nonconscious mentality. It is obvious that much of our behaviour is controlled without conscious thought. Think of driving

[3] I assume here that there is such a thing as reflective behaviour. I will defend this claim later in the chapter.

a car, playing sports, or conducting a casual conversation. When all is going well, the actions involved are spontaneous and unreflective. Indeed, giving thought to them tends to break the flow and harm performance. This unreflective mode is our default one, and nonconscious processes take care of the bulk of our behaviour.[4]

Yet this unreflective behaviour is intelligent, in the sense of being responsive to our beliefs and desires, and we would naturally explain it in belief-desire terms. For example, we would explain the way a driver manipulates the controls of their car by reference to their desires to follow a certain route and to obey the traffic laws, together with their beliefs about the workings of the controls, the rules of road, the behaviour of other road users, and so on. And we would expect their behaviour to change if these beliefs and desires changed. That is, unreflective behaviour (or much of it, at any rate) appears to be the product of practical reasoning, rationally responsive to the agent's beliefs and desires. From this perspective, the mental state underpinning Juliet's bias looks like a belief. As Schwitzgebel's description makes clear, the state affects her behaviour in a systematic way, prompting different behaviours in different contexts. Thus, we may suppose that if Juliet wanted to impress a visitor with the quality of discussion in her class, then she would avoid calling on black students to speak, but if she wanted to allow weaker students a chance to shine, then she would give black students preference. A simple association between black people and low intelligence would not affect her behaviour in this way, interacting rationally with her desires and background beliefs. If Juliet behaves like this, then, it seems, she does not merely *associate* black people with lower intelligence, she *believes* that black people have lower intelligence. The biasing state is not a rogue influence which distorts her nonconscious practical reasoning but a standard input to that reasoning.[5]

It is true that not all Juliet's behaviour fits this pattern. Some of her *reflective* behaviour (in particular, what she says) manifests a different and contradictory belief. We might think that this undermines the belief attribution and conclude that there is no clear answer as to what Juliet believes (this is the moral Schwitzgebel draws; 2010). However, there is another, and I think, more attractive, option. There are numerous dualities in folk psychology, which point to the existence of two distinct forms of belief: an implicit form, which guides thought and behaviour without being consciously recalled, and an explicit form, which requires conscious recall and affects reflective behaviour only (Frankish, 2004). I shall say more about this shortly, but, given this possibility, the conflicting evidence need not undermine the attribution of the biased belief. For the belief may be an implicit one, and the conflicting evidence,

---

[4] Even Descartes allowed that much of our behaviour is the product of nonconscious processes, including such activities as walking and singing 'when these occur without the mind attending to them', though he did not regard these processes as mental (Descartes 1642/1984, p.161).

[5] The remarks here concern everyday implicit bias like that displayed by Juliet, but there is abundant evidence that implicitly biased responses exhibited under experimental conditions are also belief-based and can be modified by argument, evidence, and logical considerations; see Mandelbaum, 2015,

which comes from Juliet's reflective behaviour, may indicate the existence of a distinct explicit one.[6]

It may be objected that the state underpinning Juliet's bias is different from the implicit beliefs and desires manifested in unreflective behaviour. For those mental states are *available* to consciousness. If a driver were to give conscious thought to what they are doing, they would draw on the same beliefs and desires that guided their unreflective behaviour. And the biasing mental state is not available to consciousness in this way. When Juliet consciously reflects on the merits of different students, she does not take it as a premise that black people are less intelligent than white people. But this in itself does not make implicit bias special. Much of the knowledge that guides our unreflective behaviour is also unavailable to consciousness. A driver might find it impossible to articulate much of the knowledge that guides their unreflective driving — about the rules of the road, the precise route they need to take, the functions of the controls, and so on. This is typical of unreflective behaviour. As William James noted, our interactions with the world are shaped by a wealth of background knowledge that we cannot articulate (using 'knowledge' in the loose sense for a body of beliefs). James cites routines such as dressing and opening cupboards: 'Our lower centres know the order of these movements, and show their knowledge by their 'surprise' if the objects are altered so as to oblige the movement to be made in a different way. But our higher thought-centres know hardly anything about the matter' (James, 1890, Vol.1, p.115). Nor, I would add, is it plausible to think that all such background knowledge was consciously acquired in the first place. Much of it is simply picked up during our routine interactions with the world. The mental states that produce implicit bias, I suggest, are of a piece with such background knowledge.

1.3 Beliefs versus attitudes

In characterizing implicit biases as grounded in beliefs, I am departing from the usual practice, which treats them as manifesting *attitudes*, in the social psychological sense, and I shall pause briefly to consider how the two characterizations differ. An attitude is an overall evaluation of something — a person, group, object, issue, and so on. Attitudes have a valence (positive or negative) and an intensity, and they are usually described as having cognitive, emotional, and behavioural aspects or components. Experimental work on attitude measurement suggests that we have two types of attitude: explicit attitudes, which are introspectable, and implicit attitudes, which are not (for surveys, see, e.g., Crano and Prislin, 2008; Nosek et al., 2011). It is this work that has stimulated recent philosophical interest in implicit bias, which is usually seen as manifesting a negative implicit attitude towards a group.

Attitudes in this sense are thus different from *propositional* attitudes, such as beliefs and desires, which are directed to propositions and have a single dimension

[6] For present purposes I assume that implicit propositional attitudes are internal representational states that play a causal role in reasoning and decision making. The overall account could, however, be modified to accommodate other views, such as dispositional ones.

(cognitive, volitional, and so on). The question of the relation between beliefs and attitudes is a complex one and turns on the precise conceptions of these states that are employed — for example, on whether attitudes are internal mental states or character traits (for the latter view, see Chapter 1.4 (Machery)). Here I shall confine myself to two general points that bear on my current strategy.

First, from a commonsense perspective at least, attitudes have beliefs as components (or as components of their bases): one's overall attitude to something is determined in part by one's beliefs about it (Webber, 2013). If we take this view, then the belief account of implicit bias and the attitude account are compatible, with the former focusing on the cognitive component of the compound state that the latter focuses on. If, on the other hand, we employ a technical notion of attitude on which attitudes do not have beliefs as components — say, one on which they are associative states of some kind — then it is doubtful that implicit biases are attitudes, for the reasons discussed in the previous subsection. Associations do not guide practical reasoning in the way that Juliet's bias does. Of course, it may be that not all implicit biases have the profile that Juliet's has, and we may need a pluralistic picture of the phenomenon to account for this.[7] But if we want to allow that *some* implicit biases are like Juliet's, then a belief account should be part of that picture.[8]

Second, the belief account of implicit bias will not be crucial to the account of self-control to be developed later in the chapter. Responding to scepticism about conscious self-control, I shall sketch a dual-level account of the mind on which conscious thought processes can, in the right circumstances, override implicit biases. This account will not depend on the belief account of implicit bias (though it will assume that there are implicit propositional attitudes), and it may be adopted by those who treat implicit bias as arising from non-propositional attitudes.

2. Dual levels

2.1 Playing double

I have argued that at least some implicit biases are the effect of implicit beliefs. However, the cases of implicit belief discussed so far did not display the dissonance that often goes with implicit bias. It is not just that Juliet does not explicitly believe that there are racial differences in intelligence, but that she explicitly believes that there are no such differences. This sort of conflict is present in many cases of implicit bias, and I shall consider it now.

There are in fact many mundane cases where a person's conscious beliefs seem to conflict with the beliefs that guide their unreflective behaviour. Consider absentmindedness (the following example is borrowed from Schwitzgebel, 2010). Ben

---

[7] On the 'heterogeneity' of implicit bias, see Holroyd and Sweetman (this volume).

[8] For other belief-based accounts of implicit attitudes, see De Houwer (forthcoming); Hughes et al., 2011; Mandelbaum, 2015; Mitchell et al., 2009. For critical discussion, see Gendler 2008a, 2008b; Levy, 2014; Madva, ms.

has been informed that a local bridge is closed and realizes he will need to take a different route to work. However, over the following days he repeatedly fails to adjust his travel plans, though he recalls the closure immediately on arriving at the bridge. Somehow, Ben's newly acquired conscious belief remains inert, and the nonconscious processes that guide his driving behaviour continue to rely on outdated information. Cases of akrasia can (perhaps surprisingly) be seen as manifesting a similar conflict. One forms the conscious intention to perform (or to refrain from) some action, but the intention remains inert and one's behaviour manifests a different intention, which has not been consciously adopted.

Of course, the dissonance in absentmindedness and akrasia is only temporary, whereas implicit bias like Juliet's is persistent. But there are everyday precedents for this too. We often accept a proposition yet fail to take it to heart and act upon it. Again, Schwitzgebel gives an example. Kaipeng is convinced by, and fully accepts, Stoic arguments for the view that death is not an evil, yet his actions and reactions are much the same as those of people who think the opposite (he fears death, regrets others' deaths, and so on). Another example comes from Mark Twain's *Huckleberry Finn*. Huck accepts the norms of his slave-owning society and believes it is wicked of him to help the escaped slave Jim. He tries to pray to change his ways, but in vain:

> [T]he words wouldn't come. Why wouldn't they? It warn't no use to try and hide it from Him. Nor from *me*, neither. I knowed very well why they wouldn't come. It was because my heart warn't right; it was because I warn't square; it was because I was playing double. I was letting *on* to give up sin, but away inside of me I was holding on to the biggest one of all. I was trying to make my mouth *say* I would do the right thing and the clean thing, and go and write to that [n-word]'s owner and tell where he was; but deep down in me I knowed it was a lie — and He knowed it. You can't pray a lie — I found that out. (Twain, 1885, p.270)

Although we would not say that Huck is implicitly *biased* (at least not if we think of bias as involving a deviation from rationality), this description of 'playing double' has strong similarities to Juliet's case. In both cases there is a conflict between the principles a person verbally endorses and what they hold onto 'deep down' — their gut instincts, manifested in their unreflective behaviour. Twain is being bitterly ironic, of course. Huck's heart is perfectly right; it is society's norms that are not. But he clearly expects his readers to find this kind of implicit/explicit conflict intelligible and recognizable.[9]

[9] For more discussion of this and other cases of 'inverse akrasia', see Arpaly, 2003; Bennett, 1974; Faraci and Shoemaker, 2014. For discussion in the context of theorizing about implicit attitudes, see Brownstein and Madva, 2012a, 2012b.

2.2 Scepticism about explicit belief

It might be suggested that the conflict in Huck's case is only apparent. Huck doesn't really believe that he ought to turn Jim in. He *says* that he should turn Jim in, and (we may suppose) *thinks* that he believes that he should do so. But his utterances reflect what he thinks he ought to say, and he is mistaken about his own beliefs. He has only one belief — the implicit belief that he should help Jim. Perhaps the same goes for Juliet too? She says she does not believe that there are racial differences in intelligence, but her behaviour shows that she is wrong about this. (The suggestion is not that she is lying about what she believes, just that she is mistaken; her self-knowledge is limited.) If we find this interpretation less plausible in Juliet's case than in Huck's, that may be simply because we regard her implicit belief, unlike his, as irrational; it is not clear that the two cases involve different types of mental state.[10] On this view, then, the dissonance in these cases is not between the subject's implicit and explicit beliefs, but between what they believe and what they think they believe.

There is in fact a strong theoretical case for endorsing this view and generalizing it. Peter Carruthers argues that (with limited exceptions) we have no direct introspective access to our own propositional attitudes — our beliefs, desires, intentions, and so on — and that our beliefs about them are the product of rapid nonconscious (and often unsound) inference from sensory evidence (Carruthers, 2011, 2014). For a mental state to be conscious, Carruthers argues, it must be globally broadcast to all cognitive systems (either because that is sufficient for consciousness or because it makes the state accessible to the mindreading system, which generates the higher-order representations required for consciousness). And the only states that are so broadcast are sensory ones. What we think of as conscious thoughts are simply sensory images in working memory, especially images of utterances (inner speech). When broadcast, these images may have important effects on our judgements, decisions, and other propositional attitudes, but they are not themselves propositional attitudes, since they do not have the right causal roles. A consequence of this is that (again with limited exceptions) there are no conscious propositional attitudes — no events of believing, desiring, judging, deciding, and so on.

On this view, then, implicit bias appears in a different light. If bias is grounded in propositional attitudes, then it is always implicit, and education and social disapproval have not driven it underground but rather changed our attitudes towards it and fostered the false belief that we are free from it. (Doubtless they have also reduced bias itself, but not by reducing explicit bias, since there is no such thing.) This view also has consequences for the control of bias. If there are no conscious decisions, then our conscious minds cannot exert control over our actions and cannot override responses arising from nonconscious processes, including biasing ones. As Carruthers puts it, 'we need to get used to the idea that most of the conscious events that we identify

[10] There is, however, evidence that people conceptualize cases like Huck's differently from ones like Juliet's, assigning praise and blame in an asymmetrical way; see Faraci and Shoemaker, 2014; Pizarro et al., 2003.

ourselves with aren't [propositional] attitudes at all, while the decisions that determine our behavior are exclusively *un*conscious in character' (Carruthers, 2011, p.380).

For present purposes, I shall grant that conscious mental events are wholly sensory in character. For I want to argue that this still leaves open a robust sense in which we can talk of having conscious beliefs and desires and making conscious judgements and decisions — and thus a sense in which we can consciously override implicit biases. In order to explain this, we need to turn to the other strand in folk psychology mentioned earlier.

2.3 Explicit belief as commitment

There are numerous dualities in folk psychology, which point to the existence of an explicit form of belief, distinct from the implicit form (see Frankish, 2004). Here I shall focus on just one, in the self-ascription of belief. We sometimes ascribe mental states to ourselves on the basis of inference from self-observation. This is common with character traits and unconscious mental states. Noticing that I place my feet oddly as I walk along the pavement, I speculate that I have a fear of treading on the cracks. Reflecting on her grading practice, Juliet infers that she has an implicit belief that there are racial differences in intelligence. Such self-ascriptions have the same status as ascriptions to other people, and we treat them as fallible. But we also self-ascribe mental states without thinking about ourselves at all. We can think simply about a state of affairs and declare an attitude towards it. Looking at photos of Hawaii, I declare that I want to go there. Reviewing the evidence for racial differences in intelligence, Juliet declares that she believes there are none. This sort of mentalistic self-ascription — *avowal* — is treated as authoritative, and a challenge to it is taken as a challenge to the speaker's sincerity or integrity. If we were to doubt Juliet's declaration that she believes that there are no racial differences in intelligence, then she would probably be affronted.

Now it could be that avowal is not really outward-looking and authoritative in the way we suppose. Perhaps it involves rapid nonconscious introspection or self-interpretation. However, there is another explanation, which justifies our intuitions about it. This is that avowals are *performative* utterances — utterances that both state that the utterer performs an action and simultaneously perform that very action. A promise, for example, is a performative utterance; in sincerely saying that I promise to do something, I make it the case that I promise to do it. If avowals are performatives, this explains their authority: a sincere avowal brings about the state it describes, and a challenge to it is a challenge to the speaker's sincerity (Frankish, 2004, ch.8; Heal 1994, 2002).

More specifically, I suggest that avowals are commitments to certain deliberate policies of reasoning and action. A key notion here is that of *acceptance* (e.g., Cohen, 1992; Engel, 2000; Frankish, 2004). To accept a proposition is to adopt a policy of *standing by* its truth — to asserting it, defending it, taking it as a premise, and acting in line with it. Acceptances can be pragmatic and context relative (lawyers accept what their clients tell them, scientists accept their hypotheses), but they can also be open-

ended and serve general epistemic ends. Explicit beliefs, I suggest, are just such open-ended acceptances. Explicit desires and intentions can be thought of as conative analogues of acceptance — policies of taking the desired outcome as a goal or of performing the intended action. (I shall use the term 'premising policies' as a general term for all these policies, cognitive and conative.) If explicit propositional attitudes are premising policies, then we can actively form them by committing ourselves to appropriate policies, and avowals, I suggest, function to self-ascribe, and thereby make, such commitments. This explains not only the authority of avowals but also why they are outward-looking. In debating whether or not to commit to standing by a certain proposition or to adopting a certain goal, we focus, not on ourselves, but on the proposition or goal itself.

I shall add two points to flesh out this suggestion. First, premising policies involve *reasoning* in certain ways. Explicitly believing that p involves taking p as a premise in one's conscious reasoning and decision-making. We can commit ourselves to doing this because conscious reasoning is — or so I claim — action-based. Within cognitive science, reasoning processes are usually thought of as subpersonal ones. But reasoning can also be an intentional, personal-level activity. We can deliberately *work things out*, motivated by (usually implicit) desires to solve problems and beliefs about what strategies may work. Strategies we can use include constructing arguments in accordance with learned rules, running thought experiments, manipulating models, diagrams, and images, and interrogating ourselves (the last serving to stimulate memory, make connections, and generate hypotheses; Dennett, 1991). These actions can be performed both overtly, in dialogue, monologue, or writing, and covertly, using inner speech or other forms of actively generated sensory imagery (see Frankish, 2004, 2009; for defence of the claim that we can actively form and manipulate sensory imagery, see Carruthers, 2009). (For convenience, I shall focus on imaged utterances from now on; similar points apply to other imagery used in conscious reasoning.)

Second, the commitments also extend to acting upon the results of this conscious reasoning. To believe something is to be guided by it in both thought and action. So if I work out that my explicit beliefs entail a certain proposition, then I am committed to adopting that proposition as a further explicit belief (or to abandoning or revising one or more of the original beliefs). And if I work out that my explicit beliefs and desires dictate that I should perform a certain action, then I am committed to performing the action (or making revisions).

On this view, overt and imaged speech acts can function as judgments and decisions. An act of saying that I believe that p (or just that p) assumes the role of the judgment that p if I treat it as a commitment to a policy of standing by p and to reasoning and acting accordingly. An act of saying that I will perform action A assumes the role of a decision to perform A if I treat it as a commitment to performing A and to planning and acting accordingly. Similarly, episodes of inner speech assume the role of occurrent beliefs and desires if we treat them as expressing premises and goals in our conscious reasoning, in accordance with prior commitments. This, in

essence, is my response to Carruthers: sensory images can assume the causal role of thoughts in virtue of being treated as such in active reasoning.

Now, treating a sensory image as a thought involves having certain propositional attitudes towards it. Treating an imaged utterance as a judgement involves (a) believing that the utterance expresses a commitment to a certain premising policy, (b) desiring to honour this commitment (or to honour such commitments generally), (c) believing that this commitment requires certain reasoning activities and overt actions (their precise nature varying with circumstances and one's other premising policies), and so on. These propositional attitudes confer the status of a judgement on the utterance and motivate the actions that are performed as a consequence of it. And these propositional attitudes will typically not themselves be explicit, conscious ones (and in the rare cases where they are, the propositional attitudes they themselves depend on will not be). Rather, they will be implicit, nonconscious states. Thus, on this view, implicit propositional attitudes partially *realize* explicit ones and make them effective in action.

What emerges, then, is a two-level picture of the human mind, with an explicit level of conscious, commitment-based states and active reasoning realized in and supported by an implicit level of nonconscious, passively formed states and involuntary processes. It is tempting to characterize these levels as *personal* and *subpersonal* (Frankish, 2009). This captures the idea that explicit reasoning is something we do, whereas implicit reasoning is done by our mental subsystems. However, it is important to add the caveat that implicit mental *states*, like explicit ones, are properly ascribed to persons.

2.4 Dual levels and dual processes

This view just outlined can be regarded as a form of dual-process theory, and since implicit bias is often discussed in the context of such theories, I shall say a little about how it differs from other theories of the type. In broad outline, dual-process theories posit two different mental processes by which a response may be generated: one (type 1) that is fast, automatic, non-conscious, and undemanding of working memory, and another (type 2) that is slow, controlled, conscious, and demanding of working memory. Type 1 processes are also variously described as associative, parallel, heavily contextualized, heuristic, and biased, and type 2 processes as rule-based, serial, decontextualized, analytical, and normative. Dual-process theories have been proposed in several fields, including deductive reasoning, decision making, social judgment, and learning and memory (for surveys, see Frankish and Evans, 2009; Frankish, 2010).[11] There are many varieties of dual-process theory and many challenges for it, including that of identifying which features are essential to each type of process and which merely typical ones.

---

[11] Some theorists have described the processes as belonging to two separate mental *systems* (e.g., Evans and Over, 1996; Stanovich, 2004). However, this description is ambiguous, and some who have used it now prefer to talk simply of two *types of processing* (Evans and Stanovich, 2013).

The view proposed here — *dual-level theory*, we might call it — can be regarded as a non-standard form of dual-process theory in which type 2 processes are identified with explicit, intentional reasoning activities involving the manipulation of sensory imagery, and type 1 processes with implicit, subpersonal reasoning processes. Many of the standard features follow from this. For example, explicit processes are slow, serial, and effortful because they involve performing sequences of actions, and they are conscious because these actions involve the manipulation of sensory imagery. Implicit processes do not involve intentional action or sensory imagery and are consequently faster, effortless, (possibly) parallel, and nonconscious. However, other standard contrasts do not carry through straightforwardly into dual-level theory. I shall mention three that are prominent in debates about implicit bias.

First, type 1 processes are typically described as automatic, and type 2 processes as controlled. Dual-level theory retains a version of this contrast: explicit processes are intentionally controlled whereas implicit ones are not. However, implicit processes are not automatic in the sense of being reflex-like, mandatory, or inflexible. As stressed earlier, much unreflective behaviour is rationally responsive to the agent's beliefs and desires and is in that sense intelligently controlled. Second, type 1 processes are often described as associative and type 2 ones as computational or ruled-governed. This contrast is not present in dual-level theory. There may be implicit associative processes of various kinds, but, as argued earlier, a great deal of implicit propositional reasoning also takes place. And although explicit thinking often involves constructing arguments in accordance with learned rules of inference, it may also involve manipulating sensory imagery in associative ways.[12] Third, type 1 processes are sometimes characterized as biased and type 2 ones as normative. On the proposed view this is only a weak contrast. It is likely that the implicit mind is modularized to some degree, and implicit belief formation and reasoning may employ a variety of heuristics and shortcuts that are adaptive but not in accord with normative theory (see, e.g., Carruthers, 2006). Explicit processes by contrast, being intentional, can be responsive to learned norms of evidence and reasoning. However, there is no general assumption that implicit processes are biased and explicit ones normative. In some contexts implicit processes may generate normative responses, and explicit reasoning and judgement may be influenced by many factors besides normative theory, including culturally acquired biases and learned rules of thumb (Carruthers, 2013b).

3. Self-control

3.1 Escaping bias

With this dual-level theory in place, I turn now to the question of how we can control implicit bias. One immediate question is how we can even *want* to control it. If we are systematically biased, how can we form unbiased judgements and motivate ourselves

[12] Compare Keith Stanovich's account of serial associative cognition (Stanovich, 2009).

to act upon them? This presents a special challenge for dual-level theory, on which explicit belief formation is motivated by implicit propositional attitudes. How does a person such as Juliet, who implicitly believes that there are racial differences in intelligence, get themselves to accept that there are no such differences?

It is true that implicit bias may impede the formation of unbiased explicit beliefs. If Juliet has an implicit belief (an intuition or gut feeling, we might say) that black people are less intelligent than white people, then this will incline her to form a corresponding explicit belief. However, she may have other implicit beliefs and desires that prompt the formation of the belief that there are no racial differences in intelligence, and these may be stronger. In particular, she may have normative beliefs about how she should think — about what counts as good evidence, the relative weight that should be given to different considerations, the untrustworthiness of gut feelings, and so on, together with beliefs about the social norms governing attitudes on this topic. And these, in conjunction with strong implicit desires to adhere to the norms in question, may induce her to accept (form the policy of premising) that there are no racial differences in intelligence, even if she still implicitly believes that there are such differences. (If social considerations play a large role, we might classify the resulting attitude as a pragmatic acceptance rather than a belief proper, but in either case, the result will be that Juliet is committed to maintaining an unbiased propositional attitude that is in tension with her implicit one.)

The mechanisms of acceptance thus offer Juliet a route by which she can escape her bias. This points to the purpose and importance of explicit cognition. By engaging in explicit reasoning and by forming and maintaining premising policies we create for ourselves a distinctively personal level of mentality, whose states and processes are available to reflection and under intentional control. The activities involved afford us a new kind of self-control, allowing us to resist responses produced by subpersonal mechanisms and to create new strategies for regulating our behaviour. Of course, this freedom is not absolute; explicit reasoning and belief formation is itself driven by implicit mental states and processes. But the explicit mind forms a new level of complexity within the overall system, shaped by normative beliefs about how one should think and modifiable in the light of tuition and reflection.

All this supposes, of course, that we can make our explicit beliefs and desires effective in action. I said that in adopting premises and goals we commit ourselves to acting upon them. But how can we do this — especially if they conflict with our implicit beliefs and desires? Suppose Juliet comes to believe that her unreflective behaviour is implicitly biased, as she might through observation of her own behaviour, the testimony of others, or inference from data about how widespread such bias is. How can she suppress her bias and ensure she is guided by her explicit belief? She might, of course, try to eradicate the implicit belief that produces the bias, but this may not be easy. Implicit beliefs cannot be formed and changed by decision. (When we talk of one-off changes of mind, we are referring to changes in our premising commitments; Frankish, 2004.) Juliet would have to employ indirect means, exposing herself to evidence and argument that undermines the implicit belief — a process

which might not succeed at all. Can she exercise a more direct form of self-control, in which her explicit belief *overrides* her implicit one? I turn to this now.

3.2 Explicit belief and action

In order to explain how explicit beliefs can override implicit ones, I need to say more about how explicit thoughts influence action. On the dual-level view an explicit thought is a self-generated (imaged) utterance, and the way in which it guides behaviour is not fundamentally different from the way in which an externally generated utterance might. In each case the influence is mediated by certain (typically implicit) propositional attitudes towards the utterance. My saying to myself that I will go to the bank does not immediately move me to go to the bank, any more than your telling me to go to the bank does. In each case I need to interpret the utterance as prompting me to go to the bank (as a commitment to going in the first case, as an instruction to go in the second) and then be motivated to act upon this prompt (desiring to fulfil my commitments or to follow instructions).

It is a consequence of this that actions guided by explicit beliefs and desires (reflective actions) will also have explanations in terms of implicit beliefs and desires. Suppose I consciously judge that I need to talk to my bank manager and consciously decide to go to the bank in the morning. Then, these explicit mental states could be cited in explanation of my subsequently going to the bank. However, the conscious decision will have become effective in virtue of implicit mental states, including a belief that I am committed to going to the bank and a desire to execute my commitments, and these implicit states could also be cited in explanation of the action. Since these implicit beliefs and desires concern my premising commitments I shall refer to them as *metacognitive* states. (Note that 'metacognitive' here does not mean *higher-order*. The implicit beliefs and desires in question are not about other implicit beliefs and desires but about the premising policies that constitute explicit beliefs and desires.)

The action thus has two different intentional explanations. This may be counterintuitive (though not more so than the idea that there are no conscious thoughts at all), but it is not incoherent or unacceptable. The two explanations are not incompatible, but pitched at different levels. If asked why a person performed a certain reflective action, we highlight the explicit beliefs and desires that prompted it. But if asked how these thoughts guided the action, we turn to lower-level processes involving implicit beliefs and desires. This is a familiar move; we give an explanation at one level, but drop down a level in order to explain the mechanisms underlying it. It is widely assumed that explicit thought processes will be susceptible to such reductive explanation in some way; the novel suggestion here is simply that the underlying mechanisms are themselves intentional (albeit involving intentional states of a different type). Even this is not unprecedented. We often highlight a certain event in the explanation of human action without mentioning the implicit beliefs and desires that make it effective. For example, we might explain why a soldier performed a certain manoeuvre by citing an order from a commanding officer. But in asking *how*

the order controlled the action, we give an explanation in terms of the soldier's implicit beliefs and desires relating to the order, their duty, the penalties for disobedience, and so on. Similarly, we might highlight the role of a promise, a warning, or a threat in the explanation of an action, and in each case, another explanation would be available that refers to largely implicit beliefs and desires about the event in question. The present proposal simply assimilates explanation in terms of conscious thoughts to this pattern.[13]

3.3 Conditions for override

With this machinery in place, we are now in a position to state conditions for the direct override of implicit bias and to understand different ways in which it may fail. Suppose subject S has an implicit belief that not-p and an explicit belief that p, in the senses outlined earlier. And suppose that in context C each of these beliefs would, in conjunction with S's other attitudes of the same type, dictate a different and incompatible action, call them A_1 and A_2 respectively. What determines whether the implicit or explicit belief guides action in C? We can highlight four necessary conditions for the explicit belief to override the implicit one.

(1) S must consciously recall p in C, representing it in inner speech or some other sensory medium. Recall is a necessary condition for override since the commitment involved in explicitly believing p is to using p as a premise *in conscious reasoning*, and conscious recall is a precondition for this. It may be asked why S could not commit simply to *acting* as if p were true, and leave the working out entirely to implicit processes. There are two points to make in reply. First, it is doubtful that the strategy would be psychologically feasible. Working out what actions the commitment requires would involve implicit *hypothetical* reasoning — bracketing one's actual implicit beliefs (which are incompatible with p) and calculating what one would do if one believed p. And there is a strong case for thinking that the capacity for hypothetical thinking depends precisely on explicit, type 2, processes (see, e.g., Carruthers, 2006; Stanovich, 2004). Second, even if the strategy were feasible, such heavy reliance on implicit processing would defeat the object of explicit belief formation. The purpose of adopting premising policies is to enhance our self-control by taking active control of our reasoning and decision making, and conscious recall of relevant inputs is required for this.

(2) S must recognize that, given their other premises and goals, p dictates that they should perform A_2. This may involve explicit argument construction, but the process

[13] Carruthers argues that if conscious mental events depend on subsequent reasoning for their efficacy, then they cannot count as decisions and judgements. A decision or judgement should settle what one will do or think, without the need for further reasoning about commitments and suchlike (Carruthers, 2011, ch.4). Again, my response is to make a distinction of levels. A conscious decision or judgement settles the matter *at the explicit level*. The further reasoning that is required to make these events effective occurs at a lower level, the level of implementation (Frankish, 2012; Carruthers responds in Carruthers, 2013a).

could be enthymematic, and the conclusion might occur to S spontaneously, courtesy of implicit processes. What is crucial is that S should realize, at least implicitly, that the conclusion is dictated by their premises and goals.

(3) S must form a commitment to performing A_2 (as opposed to revising their explicit beliefs and desires or living with inconsistency). This will often involve a conscious decision — the production of an imaged utterance that expresses a commitment to performing A_2 and is interpreted as such at an implicit level. This is not essential, however; S might implicitly realize that they are committed to performing A_2 immediately upon recalling p, without the mediation of a conscious decision. (This would be especially likely if it is obvious that A_2 is dictated and there is no temptation to do anything else.) Either way, S must form the implicit belief that they are committed to performing A_2. Assuming they also have a general desire to fulfil their commitments, they will now have implicit beliefs and desires that directly motivate A_2.

(4) S must have sufficient metacognitive motivation. Having implicit beliefs and desires that dictate A_2 does not ensure that S will perform A_2. For by hypothesis S also has implicit beliefs and desires that dictate A_1. They will perform A_2 only if their motive for doing so is stronger than their motive for performing A_1. If it is not, then S will not act on their commitment, falling into akrasia. So condition (4) is that S's desire to fulfil their premising commitments (or at least to fulfil this one) be stronger than the desire that motivates A_1 (and stronger than any other desire that motivates an incompatible action).

If these conditions are met, then, *ceteris paribus*, S's explicit belief will override their implicit one, and they will perform A_2 rather than A_1. If the conditions are not met, then override will fail. Note that these conditions do not require that S be *aware*, either consciously or nonconsciously, that they have an implicit belief that is currently prompting A_1, though they may suspect that they do and this may assist override by boosting their resolution to fulfil their premising commitments. Note, too, that the conditions can easily be modified to accommodate views on which implicit bias arises from associative attitudes rather than beliefs. We would simply revise (4) to stipulate that S's desire to fulfil their premising commitments must be strong enough to outweigh the biasing effects of the relevant implicit attitude. Thus, those who reject the belief-based account of implicit bias can still subscribe to the proposed account of self-control, provided they accept that we have implicit beliefs and desires *as well as* implicit associative attitudes.

Can we do anything to reduce the chances of override failure? We can distinguish two kinds of failure in executing a premising policy: failures of competence and failures of motivation. By failures of competence I mean failures due to lapses of memory, skill, or knowledge — for example, failure to recall a premise in a relevant context or failure to see that a group of premises dictates a certain conclusion or action (conditions (1) and (2) respectively). By failures of motivation I mean failures arising from the relative weakness of the agent's desire to execute their premising policies, or at least to execute this specific policy. The most obvious example of

motivation failure is where condition (4) is not met: a subject fails to act on their decision to perform an action because their implicit desire to execute their decision (to fulfil the commitment they have made) is, in the context, outweighed by an implicit desire to do something else. In a slightly different case, a subject might realize that their premises dictate a certain action but not be sufficiently motivated to commit to performing it, leading to a failure of condition (3). Motivation failure might also affect conditions (1)–(2). If an agent's commitment to their premises is weak, they may fail to put sufficient effort into memory search and conscious reasoning, resulting in motivational parallels to the failures of competence. (Recall itself is not, of course, under intentional control and is heavily context-dependent, but it can be intentionally stimulated by, for example, self-interrogation.) In general, high metacognitive motivation will be an important factor in effective override.

We can illustrate this by returning to Juliet and her implicit belief in intelligence differences. In some circumstances the conditions for her explicit belief to override this belief will easily be met. Suppose Juliet is asked by an academic colleague whether she thinks there are racial differences in intelligence. The question will immediately remind her of her explicit belief on the matter, and it will be immediately obvious what response it dictates, so conditions (1) and (2) will be met. And since social norms dictate the same response as her explicit belief, Juliet will feel little temptation to give a different response, even if her metacognitive motivation is relatively weak. So conditions (3) and (4) will be met too, and Juliet will say that there are no racial differences in intelligence. Contrast this with a case where Juliet is alone in her study grading essays. Her implicit belief inclines her to give lower grades to her black students, though her explicit belief dictates that she take steps to resist this and grade impartially. Here the conditions for override are less likely to hold. Since grading places a heavy load on working memory, Juliet may not recall her explicit belief at all, and if she does she may not realize what it requires of her in each case. These failures of competence may be compounded by motivational failure. If Juliet's metacognitive motivation is relatively weak, she may not make the effort required to recall and consistently apply her explicit belief, and when she does she may not have the resolution required to overcome her gut feelings about what grades different students deserve.

To sum up, in order to suppress an implicit bias, it is not sufficient to have an explicit unbiased belief and an explicit desire to be fair; one also needs a strong *implicit* metacognitive desire to make those explicit propositional attitudes effective in reasoning and action — strength of will, we might say. Failure to suppress implicit biases, I suggest, is often due to the weakness of this implicit desire.

3.4 Some predictions

I shall conclude with some predictions of the proposed account, which might form the basis for experimental work or even practical techniques for combating implicit bias. I shall not attempt to describe specific protocols, but merely sketch some ideas which others may wish to take up.

The predictions concern agents who have a biased implicit belief and an unbiased explicit belief. The general prediction is that we should be able to manipulate the relative influence of a subject's biased and unbiased beliefs in a given context by manipulating the likelihood of conditions (1)–(4) being met. Raising the chances of their being met should tend to reduce the effects of the bias, and lowering the chances of their being met should tend to increase them. (Assuming explicit beliefs can override associations as well as beliefs, this prediction should also hold if the bias arises from an associative attitude rather than a belief.)

Thus, one specific prediction is that offering reminders of the unbiased belief and its implications should reduce bias by increasing the chances of conditions (1) and (2) being met, while placing demands on working memory should increase the effects of bias by reducing the chances of those conditions being met. These predictions are, however, unlikely to be unique to the present account (though their falsity would of course undermine it).

Other specific predictions focus on motivation. Since override is motivated, we should be able to manipulate it by manipulating the agent's motivational state. Thus, boosting an agent's motivation to execute the premising policy that constitutes the unbiased belief should increase the chances of all four conditions being met and so reduce the effect of the bias, whereas reducing this motivation should have the opposite effect.[14] Boosting might be achieved by providing subjects with direct or indirect reminders of the importance of the issue and of the harmful effects of the bias, and reduction by offering contrary suggestions. Similar effects should be obtainable by manipulating the agent's motivation for acting on their implicit belief. For example, if a subject is told that their interlocutor knows their 'gut feelings' and will reward them for acting on them, then this should boost their motivation for acting on their implicit belief, reducing the likelihood of override.[15] (An extreme version of this scenario is talking to God. As Huck says, you can't pray a lie.) Again, however, these predictions are unlikely to be specific to the present account; many theories will predict correlations between a subject's attitudes towards their bias and the likelihood of their acting on it.

Perhaps the most distinctive prediction of the account is that we can manipulate implicit bias by manipulating a subject's desire to honour their commitments *generally*. Boosting this general desire should have the knock-on effect of boosting their specific desire to execute their premising commitments and so reduce the effects of implicit bias. This might be achieved by priming subjects with suggestions of the importance of keeping promises and sticking to commitments, or presenting them with stimuli associated with integrity, consistency, self-discipline, and strength of will

---

[14] For potentially relevant empirical findings, see studies on 'implicit motivation to control prejudice' (e.g., Glaser and Knowles, 2008; Park et el., 2008; Park and Glaser, 2011).

[15] See 'bogus pipeline' manipulations (e.g., Nier, 2005) in which subjects are told that the Implicit Association Test (used to measure implicit attitudes) is akin to a lie-detector test. The result is a closer correlation of subjects' implicit and explicit attitudes.

(controlling of course for possible confounding factors). Suggestions and stimuli that tend to weaken a subject's desire to keep their promises and honour their commitments should have the opposite effect and increase the effects of bias. These predictions arise from the role commitment plays in the dual-level account of implicit bias, and as far as I know they are unique to the account. Confirmation of them would therefore be strong support for it.

Conclusion

On the view I have sketched implicit bias is more a part of us than we may like to think, and perhaps more natural to us too, reflecting the operations of subpersonal belief-forming mechanisms that were designed to be adaptive not impartial. We all play double sometimes. But while bias may be natural, so is the capacity to overcome it. Our ability to engage in explicit thought is one of our most distinctively human features, and with sufficient strength of will we can use it to reflectively control our actions, override our biases, and become better, fairer people.[16]

Works Cited

Arpaly, N. (2003). *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford: Oxford University Press.

Bennett, J. (1974). The conscience of Huckleberry Finn. *Philosophy*, 49(188), 123-34.

Bosson, J. K., Swann, W. B. Jr., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: the blind men and the elephant revisited? *Journal of Personality and Social Psychology*, 79(4), 631-43.

Brownstein, M. and Madva, A. (2012a). Ethical automaticity. *Philosophy of the Social Sciences*, 42(1), 68-98.

Brownstein, M. and Madva, A. (2012b). The normativity of automaticity. *Mind & Language*, 27(4), 410-34.

Carruthers, P. (2006). *The Architecture of the Mind: Massive Modularity and the Flexibility of Thought*. Oxford: Oxford University Press.

Carruthers, P. (2009). An architecture for dual reasoning. In J. St. B. T. Evans and K. Frankish (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 109-27). Oxford: Oxford University Press.

Carruthers, P. (2011). *The Opacity of Mind: An Integrative Theory of Self-Knowledge*. New York, NY: Oxford University Press.

Carruthers, P. (2013a). On knowing your own beliefs: a representationalist account. In N. Nottelmann (ed.), *New Essays on Belief: Constitution, Content and Structure* (pp. 145-65). Basingstoke: Palgrave Macmillan.

[16] Thanks for comments and advice are due to Michael Brownstein, Ward Jones, Maria Kasmirli, Jenny Saul, an anonymous referee, and the participants in the Implicit Bias and Philosophy project.

Carruthers, P. (2013b). The fragmentation of reasoning. In P. Quintanilla (ed.), *La Coevolución de Mente y Lenguaje: Ontogénesis y Filogénesis* (pp. 181-204). Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú.

Carruthers, P. (2014). On central cognition. *Philosophical Studies*, 170(1), 143-62.

Cohen, L. J. (1992). *An Essay on Belief and Acceptance*. Oxford: Oxford University Press.

Crano, W. D. and Prislin, R. (eds.). (2008). *Attitudes and Attitude Change*. New York, NY: Psychology Press.

De Houwer, J. (forthcoming). A propositional model of implicit evaluation. *Social Psychology and Personality Compass*.

Dennett, D. C. (1991). *Consciousness Explained*. New York: Little, Brown and Co.

Descartes, R. (1642/1984). Objections and replies. In J. Cottingham, R. Stoothoff, and D. Murdoch (eds.), *The Philosophical Writings of Descartes: Volume 2* (pp. 63-383). Cambridge: Cambridge University Press.

Engel, P. (ed.). (2000). *Believing and Accepting*. Dordrecht: Kluwer.

Evans, J. St. B. T. and Over, D. E. (1996). *Rationality and Reasoning*. Hove: Psychology Press.

Evans, J. St. B. T. and Stanovich, K. E. (2013). Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-41.

Faraci, D. and Shoemaker, D. (2014). Huck vs. JoJo: moral ignorance and the (a)symmetry of praise and blame. In J. Knobe, T. Lombrozo, and S. Nichols (eds.), *Oxford Studies in Experimental Philosophy: Volume 1* (pp. 7-27). Oxford: Oxford University Press.

Frankish, K. (2004). *Mind and Supermind*. Cambridge: Cambridge University Press.

Frankish, K. (2009). Systems and levels: dual-system theories and the personal-subpersonal distinction. In J. St. B. T. Evans and K. Frankish (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 89-107). Oxford: Oxford University Press.

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass*, 5(10), 914-26.

Frankish, K. (2012). Dual systems and dual attitudes. *Mind & Society*, 11(1), 41-51.

Frankish, K. and Evans, J. St. B. T. (2009). The duality of mind: an historical perspective. In J. St. B. T. Evans and K. Frankish (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 1-29). Oxford: Oxford University Press.

Gawronski, B., Hofmann, W., and Wilbur, C. J. (2006). Are "implicit" attitudes unconscious? *Consciousness and Cognition*, 15, 485-99.

Gendler, T. S. (2008a). Alief and belief. *The Journal of Philosophy*, 105(10), 634-63.

Gendler, T. S. (2008b). Alief in action (and reaction). *Mind & Language*, 23(5), 552-85.

Glaser, J. and Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology*, 44(1), 164-72.

Hahn, A., Judd, C. M., Hirsh, H. K., and Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3), 1369-92.

Heal, J. (1994). Moore's paradox: a Wittgensteinian approach. *Mind*, 103(409), 5-24.

Heal, J. (2002). On first-person authority. *Proceedings of the Aristotelian Society*, *102*, 1-19.

Holroyd, J. and Sweetman, J. This volume. The Heterogeneity of Implicit Bias.

Hughes, S., Barnes-Holmes, D., and De Houwer, J. (2011). The dominance of associative theorizing in implicit attitude research: propositional and behavioral alternatives. *The Psychological Record*, 61, 465-96.

James, W. (1890). *The Principles of Psychology*. New York: Henry Holt and Co.

Levy, N. (2014). Neither fish nor fowl: implicit attitudes as patchy endorsements. *Noûs*. doi: 10.1111/nous.12074

Madva, A. (ms). Why implicit attitudes are (probably) not beliefs.

Mandelbaum, E. (2015). Attitude, inference, association: on the propositional structure of implicit bias. *Noûs*. doi: 10.1111/nous.12089

Mitchell, C. J., De Houwer, J., and Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, 32, 183-246.

Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations*, 8(1), 39-52.

Nosek, B. A., Hawkins, C. B., and Frazier, R. S. (2011). Implicit social cognition: from measures to mechanisms. *Trends in Cognitive Sciences*, 15(4), 152-9.

Olson, M. A. and Fazio, R. H. (2003). Relations between implicit measures of prejudice: what are we measuring? *Psychological Science*, 14(6), 636-9.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology*, 105(2), 171-92.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: small effect sizes of unknown societal significance. *Journal of Personality and Social Psychology*, 108(4), 562-71.

Park, S. H. and Glaser, J. (2011). Implicit motivation to control prejudice and exposure to counterstereotypic instances reduce spontaneous discriminatory behavior. *Korean Journal of Social and Personality Psychology*, 25(4), 107-20.

Park, S. H., Glaser, J., and Knowles, E. D. (2008). Implicit motivation to control prejudice moderates the effect of cognitive depletion on unintended discrimination. *Social Cognition*, 26(4), 401-19.

Pizarro, D., Uhlmann, E., and Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychological Science*, 14(3), 267-72.

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.

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago: University of Chicago Press.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: is it time for a tri-process theory? In J. St. B. T. Evans and K. Frankish (eds.), *In Two Minds: Dual Processes and Beyond* (pp. 55-88). Oxford: Oxford University Press.

Twain, M. (1885). *Adventures of Huckleberry Finn (Tom Sawyer's comrade)*. New York: C. L. Webster.

Webber, J. (2013). Character, attitude and disposition. *European Journal of Philosophy*. doi: 10.1111/ejop.12028.