Philos Stud (2019) 176:2961–2990
https://doi.org/10.1007/s11098-018-1159-7

Implicit attitudes and the ability argument

Wesley Buckwalter[1]

Published online: 15 September 2018

Abstract According to one picture of the mind, decisions and actions are largely the result of automatic cognitive processing beyond our ability to control. This picture is in tension with a foundational principle in ethics that moral responsibility for behavior requires the ability to control it. The discovery of implicit attitudes contributes to this tension. According to the ability argument against moral responsibility, if we cannot control implicit attitudes, and implicit attitudes cause behavior, then we cannot be morally responsible for that behavior. The purpose of this paper is to refute the ability argument. Drawing on both scientific evidence in cognitive science and philosophical arguments in ethics and action theory, I argue that it is invalid and unsound because current evidence is insufficient to establish the premises that (1) implicit attitudes are uncontrollable, (2) that they significantly cause behavior, (3) that responsibility always requires ability, and (4) that even if uncontrollable attitudes did fully cause behavior, this entails that the behavior they cause is uncontrollable. The rejection of the ability argument questions the priority of the unconscious over the conscious mind in cognitive science, deprioritizes ability in theories of moral responsibility in ethics, and provides a strong reason to uphold moral responsibility for implicitly biased behavior.

Keywords Implicit attitudes · Bias · Ability · Choice · Control · Belief · Ought implies can

✉ Wesley Buckwalter
wesleybuckwalter@gmail.com

[1] Department of Philosophy, School of Social Sciences, University of Manchester, Oxford Road, Manchester M13 9PL, USA

1 Introduction

According to one conception of the mind popular for over three decades, cognitive processes can be grouped into two basic systems (Evans 2003; Moskowitz et al. 1999; Sloman 1996). One system contains "lower level" automatic processes that reside below full conscious awareness or ability to control. The other set of processes are "higher level" executive cognitive processes. These are conscious processes within our ability to intentionally control. This basic dual-process understanding of human cognition raises a fundamental question at the intersection of philosophy and psychology: which system has priority over our beliefs, evaluations, decisions, and ultimately, how we live our lives?

On one answer, the automatic takes priority (Bargh 1999; Bargh and Chartrand 1999; Bargh and Williams 2006). According to this picture of the mind, "most of a person's everyday life is determined not by their conscious intentions and deliberate choices but by mental processes that are put into motion by features of the environment and that operate outside of conscious awareness and guidance" (Bargh and Chartrand 1999: 462). Other researchers have argued that automatic processes "often act as key levers in complex systems that give rise to social problems" and that while "social problems are complex and multicaused…nearly every problem involves psychology" (Walton 2014: 80). Supporting this view is a mountain of social psychological claims on the influence of automatic processes over beliefs and behavior of moral and social significance. For example, heuristics and biases (Kahneman 2011; Tversky and Kahneman 1974), implicit procedural learning (Cleeremans and Jimenez 2002), unconscious social priming (Payne et al. 2016), subtle wording effects on voter turnout (Bryan et al. 2011), growth mindset (Blackwell et al. 2007), and stereotype threat (Steele and Aronson 1995) all appear to support, to one degree or other, what Bargh and Chartrand call the "unbearable automaticity of being".

Some of the most well researched automatic processes in social psychology are implicit attitudes (Greenwald and Banaji 1995; Nosek et al. 2007a). This is the discovery that we automatically form unconscious attitudes towards groups, people, or objects that can sometimes conflict with our conscious attitudes and considered judgments toward them. One important focus of research in implicit social cognition studies implicit attitudes involving socially and morally significant categories. For example, while many might explicitly report the egalitarian belief that they have no preference for whites over African Americans, they are also likely to be implicitly biased against African Americans (e.g. Nosek et al. 2007b). Similar implicit biases have been demonstrated against women (Dasgupta and Asgari 2004), Muslims (Park et al. 2007), the elderly (Castelli et al. 2005), the obese (O'Brien et al. 2007; Teachman et al. 2003), and persons with mental illness (Rusch et al. 2010; Teachman et al. 2006).

The existence of implicit attitudes involving significant social and moral categories such as race, gender, religion, or disability has motivated a large body of research on the role that implicit attitudes might play in explaining persistent discriminatory behavior and unjust social outcomes. Subsequent research has investigated the possible link between implicit attitudes and employment discrimination (Bendick et al. 1994; Moss-Racusin et al. 2012), housing practices (Ahmed and Hammarstedt 2008), health care disparities (Sabin et al. 2009), and policing (Correll et al. 2002; Payne 2006; Sim et al. 2013). For example, researchers studying implicit bias in policing find evidence for implicit attitudes associating race with the presence of weapons, and that this sometimes predicts results in laboratory first-person shooter video game experiments (Glaser and Knowles 2008). Researchers study these connections, in part, to uncover whether implicit attitudes could predict or explain real world cases of mistaken shootings.

The discovery of implicit social cognition has also dominated philosophical inquiry in ethics, action theory, and philosophy of mind (Fricker 2007; Levy 2015; Machery 2016; Mandelbaum 2016; Saul 2012; Schwitzgebel 2010). One important focus in this line of research also involves the possibility that implicit attitudes predict or explain our beliefs and real-world behavior. For example, one central question in this line of research is whether agents can be morally responsible for discriminatory behaviors when those behaviors are caused by implicit attitudes (Brownstein 2016; Brownstein and Saul 2016; Holroyd 2012; Kelly and Roedder 2008; King and Carruthers 2012; Levy 2014, 2017; Washington and Kelly 2016). Is an agent morally responsible for a mistaken shooting, for instance, if this action was caused by attitudes that are beyond their ability to control?

According to one foundational principle in ethics, the answer is "no". This is the principle that having a moral responsibility for an action at a certain time requires that an agent have the ability to control that action at that time. Versions of this principle weave their way through the history of western philosophy and continue to loom large in present day discussions of responsibility. For example, the principle is shared by the majority of philosophers who accept that ought implies can (Cicero and Edmonds 1856; Copp 2008; Feldman 1986; Hare 1965; Kant 1998; Moore 1922; Van Fraassen 1973; Vranas 2007). Versions of this principle are held by theorists who believe that moral responsibility requires the ability to do otherwise (Blum 2000; Copp 2008) and is invoked in discussions of determinism, incompatiblism, and free will (Fischer 2003; van Inwagen 1983; Widerker 1991; Woolfolk et al. 2006). According to one theorist, the claim that "an agent is morally responsible for an action or for the consequences of an action only if she exercised 'freedom-level' control over that action or that consequence" is a condition on responsibility that "almost every prominent theorist accepts," in some form or another (Levy 2017: 5).

The ability condition stands in tension with the view that the automatic takes priority in human cognition. If most of our actions are not determined by deliberate choices, then it is possible that we regularly lack control over our actions, which according to the ability condition, would undermine moral responsibility for them. The discovery of implicit biases in cognitive science and their possible association with discriminatory behavior makes this tension with a foundational principle of ethical theorizing pointed and concrete. This tension is perhaps best articulated by Neil Levy, who argues that the control we have over our actions is "greatly diminished" by the existence of implicit attitudes and that "the decrease in control is significant enough to make it highly plausible that the agent lacks responsibility-level control" for certain actions and consequences that occur as a result of them (2017: 6).

Applying this foundational condition on moral responsibility in ethics to implicit bias research, the condition motivates the following argument against moral responsibility (for other versions of this argument see Brownstein 2017; Holroyd 2012). Call this the *ability argument* against implicit attitudes:

1. S does not have the ability to control implicit attitude *p*.
2. Implicit attitude *p* causes action ϕ.
3. If S is morally responsible for ϕ, then S has the ability to control ϕ.
4. Therefore S cannot be morally responsible for ϕ.

If the conclusion of the ability argument against moral responsibility were correct, this would rule out moral responsibility for actions caused by implicit attitudes, such as the case of the mistaken shooting or other police misconduct, when it follows as the direct result of implicit bias.

The purpose of the paper is to reject the ability argument against moral responsibility. To do this I draw on both scientific evidence and philosophical arguments concerning implicit attitudes, agency, and moral responsibility. First, I argue that the ability argument is unsound. Current evidence does not establish its premises, and at present writing, at least, it appears likely that each premise of the ability argument is false. Instead, it is likely that implicit attitudes are controllable, that they do not significantly cause behavior, and that responsibility may not always require ability. Second, and perhaps more importantly, the ability argument against moral responsibility for implicitly biased behavior is invalid. Granting these premises does not rule out moral responsibility for behavior caused by implicit attitudes because even if uncontrollable attitudes cause a behavior, this does not entail that a behavior they cause is uncontrollable. These challenges to the ability argument question the priority of the unconscious over the conscious mind in cognitive science, deprioritize ability in theories of moral responsibility in ethics, and provide a strong reason to uphold moral responsibility for implicitly biased behavior of important social concern.

Before proceeding, it is important to empathize that these assessments should not be viewed as settling the question regarding moral responsibility for actions caused by implicit attitudes rather than evaluating a specific argument against it given a current state of evidence. Implicit social cognition is a dynamic area of social scientific and philosophical research in which new evidence is rapidly being discovered and disseminated. It is also a research area where consensus has not been reached on some key issues pertaining to the structure, impact, and content of implicit attitudes. Thus, the best way to make progress on philosophical questions pertaining to implicit attitudes is to charitably evaluate ongoing research as discoveries are made or questioned and to update beliefs accordingly. In light of this, the present contribution might best be perceived as identifying what has not been sufficiently demonstrated to make the ability argument for moral responsibility go, isolating the specific things that need to be shown in the future to make that argument go, and invite future philosophical and scientific exploration of those things as research in this area continues to progress.

2 Control of implicit attitudes

The first premise of the ability argument is that we do not have the ability to control implicit attitudes. However, this premise is not sufficiently demonstrated and current evidence appears to point in the opposite direction. Some early characterizations of implicit attitudes presumed that implicit attitudes are inflexible (Bargh 1999; Wilson et al. 2000). But subsequent research suggests that implicit attitudes are malleable, to various extents (Blair 2002; Chapman et al. 2018; Dasgupta and Greenwald 2001; Forscher et al. 2018; Frankish 2016; Lai et al. 2013; Lenton et al. 2009; Olson and Fazio 2006). The most recent and perhaps most compelling evidence for this comes from a meta-analysis examining different procedures for changing several kinds of states and actions, including implicit attitudes, implicit stereotypes, explicit beliefs, and behavior (Forscher et al. 2018). The analysis represented over eighty thousand participants across over three hundred articles published over the last 20 years. It included between-subjects experiments measuring implicit attitude change using assorted measures of implicit bias (e.g. the implicit association test, lexical decision tasks, affect misattribution procedures, go/no-go association tasks), that associated a variety of attributes (e.g. good/bad, presence of stereotype) toward a wide range of implicit targets (e.g. whites, the elderly). The analysis also included studies incorporating a diverse range of experimental procedures for changing implicit bias (e.g. association priming, evaluative conditioning). The principle finding of this research was to confirm that implicit attitudes can change and to identify which procedures were effective at changing them over others that were not effective.

One set of experimental procedures that were found to change implicit attitudes most effectively were classified as attempts to "strengthen or weaken implicit associations directly". Included in this group were a number of different experimental manipulations such as counter-stereotypical exemplars (Dasgupta and Greenwald 2001), deliberative information processing (Horcajo et al. 2010), and evaluative conditioning (Olson and Fazio 2006). In one set of studies involving deliberative information processing, for example, researchers showed that some implicit attitudes are responsive to rhetorically persuasive arguments (Horcajo et al. 2010). In one experiment, the researchers demonstrated that participants showed more positive implicit attitudes toward vegetable consumption when asked to carefully consider the argument that "vegetables have more vitamins than most supplements on the market, making them particularly beneficial during exam and workout periods" than those presented with advertisements for a neutral topic (Experiment 1, 943). The researchers also demonstrated that explicit arguments in favor of one concept could impact implicit attitudes toward other concepts related to it. For example, researchers demonstrated that participants who saw arguments advocating for the color green (used in a University logo) also had more positive implicit associations with the brand Heineken (a brand associated with the color green) than those presented with arguments not advocating for green (2010: Experiment 2). The researchers take these findings as evidence that argumentation and persuasion can change implicit attitudes.

Other studies included in this category include mitigating bias through exposure to counter-stereotypical exemplars and intergroup contact (Blair et al. 2001; Dasgupta and Greenwald 2001; Dasgupta and Rivera 2008; Gonsalkorale et al. 2010; Ramasubramanian 2011; Turner and Crisp 2010). In one classic paper, for example, researchers demonstrated that implicit attitudes (though not explicit attitudes) against African Americans and the elderly were reduced when participants were presented with pictures of admired African Americans (e.g. Denzel Washington) or old (e.g. Mother Teresa) individuals (Dasgupta and Greenwald 2001). Other researchers have replicated and expanded these findings by demonstrating that reduction of negative implicit attitudes occurs by simply imagining intergroup contact with members of certain groups, for example, the elderly and Muslims (Turner and Crisp 2010), and immigrants (Vezzali et al. 2011). This research indicates that implicit bias is malleable and can be reduced through pictures of counterstereotypical exemplars or imagined contact with group members.

Another category of effective experimental procedures identified by Forscher et al. (2018) involved directly or indirectly invoking goals to strengthen or weaken implicit attitudes (for a review, also see Lai et al. 2013). Included in this group were a number of different experimental manipulations involving the use of norm or value priming (Blincoe and Harris 2009; Jonas et al. 2010), implementation intentions (Mendoza et al. 2010; Stewart and Payne 2008), motivation priming (Legault et al. 2011), and training (Plant et al. 2005). For example, researchers found that participants were able to reduce stereotypes in a weapon sorting task when they adopted the goal to respond as accurately as possible by intentionally thinking the word “safe” whenever they saw a black face (Stewart and Payne 2008: Experiment 2). Other researchers found that implicit attitudes were impacted by simple directives (Wallaert et al. 2010). More specifically, these researchers found that participants showed greater pro-white bias on a race IAT when instructed to respond like “someone who holds a strong preference for Whites over Blacks” and less pro-white bias after they were directly told to “please be careful not to stereotype on the next section of the test” (2010: 4–5).

Other researchers have expanded upon these effects by demonstrating that the efficacy of directives to promote egalitarian goals may depend on the kind of motivations they invoke (Legault et al. 2011). More specifically, researchers found that participants primed with self-determined motivation to reduce prejudice (e.g. “I can freely decide to be a nonprejudiced person”) showed significantly less implicit bias against blacks than those primed with motivations that involved social control (e.g. “It is socially unacceptable to discriminate based on cultural background”). In fact, use of the controlling prime resulted in higher implicit bias scores against blacks than doing nothing, while participants who saw the self-determined prime displayed no preference for white or blacks in the experiment at all (Legault et al. 2011: Experiment 2). The researchers concluded that implicit biases are malleable and that emphasizing personal autonomy may actually be crucial for reducing them.

These results support the conclusion that implicit attitudes are changeable, with four important caveats. First, it is important not to exaggerate the effect size of implicit attitude change. While meta-analysis indicated that the attitudes were malleable, the effect size was small. Second, this research did not show how long the changes to implicit attitudes by these experimental procedures persisted, and subsequent research suggests the duration may be short (Lai et al. 2016).[1] Third, biases can also sometimes change as a result of situational factors, for example, those that lead to extraneous cognitive load (Allen et al. 2009). Fourth, most studies were conducted in laboratory settings among university students, which might question their generalizability to other domains or samples. These caveats notwithstanding, however, research to date strongly suggests that implicit attitudes are malleable in principle, and that some of the most effective procedures for changing them emphasize the agency, goals, intentions, and motivations of individuals.

Supposing that implicit biases are changeable, a further philosophical question is whether "changeable" means "controllable", in the sense that is denied by the first premise of the ability argument. The answer to this question will likely depend on substantive accounts of control and distinctions between types of responsibility and control that one accepts. According to at least one such distinction, there are two important senses of "responsibility" relevant to these discussions, namely, "indirect" and "direct" responsibility (see Arpaly 2003; Holroyd 2012; Levy 2017). The difference between them involves the kind of control we have over the behavior in question. On one account of this difference, an implicitly biased agent has an indirect responsibility for their behavior at time *t* when it is "reasonable to expect her to try to change her implicit attitudes prior to *t*" but lacks "direct" responsibility when it would not be reasonable to expect her to control her behavior at *t* (Levy 2017: 4). Many agree that implicitly biased agents have indirect responsibilities for their behavior. For example, it is reasonable for implicitly biased agents to control future discriminatory behavior by adopting institutional policies and procedures widely known to mitigate biased outcomes. Anonymous grading, employee evaluation, or resume searchers, for instance, could prevent biases from influencing evaluations. The techniques researched above continue to suggest that control of attitudes is consistent with this type of responsibility.

However research also strongly questions whether the kind of control we have over implicit attitudes rules out "direct" responsibility for implicitly biased behavior. Whether or not the ability to change an implicit attitude is necessary for control of a behavior is the topic of subsequent sections of this paper. The present question is whether it has been demonstrated that we lack a kind of immediate control over implicit states at the time of an action, such that, given what an agent's implicit attitudes are at *t*, it is reasonable to expect her to control her behavior at *t*, even if this does require changing implicit attitudes. Here again, however, research has not ruled out this possibility and even suggests that it may be possible. It appears

[1] Also pertaining to this issue are related research questions concerning the temporal stability, test–retest reliability, and measurement of implicit attitudes (Cooley and Payne 2017; Gawronski et al. 2017).

that some of the most effective mechanisms for implicit attitude change involve emphasizing agency, and perhaps even a kind of direct or conscious engagement with attitudes, goals, intentions, and reasons. For example, to the extent that strategies like thinking the word “safe” to oneself or contemplating one’s goals about safety when trying to avoid harming someone are successful, these mechanisms are likely to be part of the underlying processes that are associated with and invoked with those when one is typically trying not to harm someone. If the nature of several successful techniques reviewed above have this character, emphasizing agency and associated underlying processes with goals and intention, then it suggests that the amount of control one ultimately has over these states does not make it unreasonable to expect implicitly biased agents to control attitudes and actions of social and ethical significance.

This growing body of evidence also suggests that, in some cases at least, implicit attitudes are more easily changed than explicit beliefs (Dasgupta and Greenwald 2001; Forscher et al. 2018). For example, in addition to implicit measures, Forscher et al. (2018) also studied the impact of experimental procedures on several explicit measures (e.g. modern racism scale, self-reports). The researchers found that effects on explicit biases were smaller overall and that effect sizes for implicit attitude change were significantly greater than explicit attitude change for the procedures that weakened goals and associations. This comparison usefully helps situate the question of implicit attitude control within discussions of mental state control in philosophy of mind. By way of analogy, a foundational research question is whether explicit beliefs are voluntarily controllable at will. Most contemporary philosophers agree that beliefs are not voluntarily controllable, at least directly (see, for example Alston 1988; Bennett 1990; Williams 1973). The research above suggests that implicit attitudes are sometimes subject to direct control, for example, when implicit attitudes change in response to instructions, motivations, or goals. But if implicit biases are just as malleable as explicit biases, or perhaps even more so using some procedures, then it is unclear how the issue of controllability of implicit attitudes creates a distinct problem in ethics over and above the question of moral responsibility for behavior caused by explicit states such as explicit prejudicial beliefs.

To summarize, there is insufficient evidence to support the premise that implicit attitudes are not controllable in at least two important senses of control. Instead, evidence to date suggests that they may be controllable, undermining this premise in the ability argument, with two important caveats. First, research about the controllability of implicit states is ongoing and research to date has largely been conducted in the lab, online, and other experimental settings. These experiments demonstrate how implicit attitudes can be changed not whether people in the real world do in fact change them. It is an open question whether or how often these effects, if reliable, practically extend to real world situations at the time of a behavior. At the same time, of course, the same is true of a significant portion of implicit attitude research concerning the presence of an implicit attitudes in the first place. Second, it is possible that implicit attitudes are controllable but that it is nonetheless more or less difficult to control them in some situations over others, just as it sometimes more or less difficult to control various physical actions in different situations. Further research might explore how this related but distinct issue of difficulty, rather than controllability of behavior causing attitudes could potentially be developed into a separate argument for reducing the degree of moral responsibility that is present for those actions.

3 Behavior and causation

Understanding how implicit attitudes change is crucial for furthering our understanding of implicit social cognition in psychology and cognitive science. The research is also motivated by the promise that identifying and changing implicit attitudes may promote positive social outcomes, for example, by reducing prejudicial or discriminatory behavior. This possibility assumes that there is a causal link between individual differences in implicit attitudes and behavior whereby changing implicit attitudes will translate into changes in real-world behavior. While this is an open empirical possibility that merits continued research, the current evidence for this claim is mixed. Some evidence speaks against the claim that implicit attitudes are causal, while other evidence is at least consistent with the claim they are causal, by demonstrating that implicit attitudes do make contributions to predicting behavior. This section reviews these sources of evidence for implicit attitudes as understood through measures of individual differences and evaluates whether they are sufficient for accepting the causal premise of the ability argument.

The first source of evidence comes from meta-analyses that either find modest predictive relationships or no relationships between implicit attitudes and person-level behavioral outcomes (Carlsson and Agerstrom 2016; Correll et al. 2014; Greenwald et al. 2009; Mitchell 2018; Oswald et al. 2013, 2015). There is disagreement between researchers stemming from the use of different inclusion criteria and bias correction procedures between meta-analyses about the correct size of this association. Some researchers estimate the effect as small, ranging from d = .14 to .24 (Greenwald et al. 2009; Oswald et al. 2013) while others argue that by using more fine-grained criteria for what counts as discriminatory behavior, the predictive relationship is or is near zero (Carlsson and Agerstrom 2016). These ongoing debates notwithstanding, there is agreement that implicit attitudes are likely to be only modestly associated with individual behavior.

A second source of evidence of association comes from studies that investigate the relationship between implicit attitudes and specific actions of social concern, with mixed results. For example, researchers have found that although implicit attitudes do impact performance in laboratory first-person shooter simulations, implicit attitudes are better predictors of certain individual behaviors over others. This is shown in a recent meta-analysis of forty-two shooter task studies of racial prejudice (Mekawi and Bresin 2015). The researchers confirmed the existence of racial bias in individual shooter task performance. Specifically, the research showed that participants do shoot simulated black targets more quickly and more often than white targets when those targets were armed (i.e. did predict non-mistaken shootings). However, the research also found that participants do not shoot simulated black targets more than white targets when the targets were unarmed (i.e.

did not predict mistaken shootings). Similar results were found by other researchers studying shooter bias among law enforcement professionals. Specifically, these researchers found that although race does impact response times in shooter simulations, the officers were no more likely to mistakenly shoot unarmed black or white targets during shooter simulations (Correll et al. 2014, 2007). This evidence is mixed, since it is consistent with the hypothesis that implicit states may predict some individual behaviors, such as non-mistaken shootings, on the one hand, but questions the hypothesis that it predicts other kinds of behaviors of ethical concern, such as mistaken shootings, on the other. Alternatively, subsequent research suggests that implicit attitudes may be predictive of this on the group or aggregate level (Hehman et al. 2018; see also Payne et al. 2017). Researchers have found that regional implicit attitudes do correlate with crowd-sourced reports of lethal force against African Americans disproportionally to regional population rates. Of course, there are many possible interpretations of these results. However, one interpretation is that implicit attitudes may correlate with disparate outcomes on the aggregate or societal level without necessarily needing to posit significant person-level associations to explain those outcomes. For example, they might arise due to situational factors rather than individual differences.

A third source of evidence comes from studies challenging the idea that implicit attitudes make unique contributions to predicting behavior over and above other related factors. Researchers have found evidence that some explicit measures are more strongly correlated with behavioral outcomes than IAT scores are across several important social domains (Greenwald et al. 2009; Oswald et al. 2013). For example, in a large meta-analysis, researchers found that implicit attitudes measured through IATs were no better predictors of policy preferences, interpersonal behavior, person perceptions, reaction times or misbehavior (Oswald et al. 2013: 183). Other researchers have questioned whether implicit attitudes aren't actually consciously accessible after all (Hahn et al. 2014). To that end, these researchers demonstrate that participants predict their own implicit attitudes with a high degree of accuracy. Relatedly, and contrary to how it may at first appear, other researchers have questioned whether implicit attitudes and explicit attitudes don't actually tap into the same underlying construct or whether certain implicit measures have been shown to provide evidence of a distinctly unconscious process (Hofmann et al. 2005; Schimmac 2017). These results suggest that the correlation between implicit attitudes and discriminatory behavior could largely be explained by explicit attitudes, without necessarily needing to posit a causal relationship with implicit attitudes to explain that behavior.

Contrary to this, however, a subsequent meta-analysis argues that implicit attitudes do correlate with criterion measures of intergroup behavior after controlling for explicit attitudes (Kurdi et al. forthcoming). In an analysis of 217 studies, researchers demonstrate that implicit and explicit measures each make unique contributions to predicting behavior ($\beta = .14$ and .11, respectively) across various domains (with a prediction interval ranging from $r = -.14$ to .32, across domain of intergroup discrimination). This is encouraging evidence that implicit attitudes are uniquely associated with behavior on par with explicit states, with two caveats. First, the study uses inclusive criteria for intergroup behavior that include items that are not typically the focus of ethical evaluation or discussions of discriminatory behavior, such as physiological indicators. A better assessment of the link to discriminatory behavior would exclude these from the analysis. Second, given prior findings on the relationship between implicit and explicit states, it is possible that the development of better or different explicit measures could account for more of the variance than is presently accounted.

Lastly, a fourth source of evidence comes from Forscher et al. (2018)'s meta-analysis of implicit bias change. In addition to analyzing the effect of experimental procedures on implicit and explicit attitude change, the researchers also investigated their impact on behavior and the relationship between these things. Several measures of behavior were used across studies included in the meta-analysis, including for example, budget allocation (Yoshida et al. 2012), resume rating (Gapinski et al. 2006), or seating distance (Mann and Kawakami 2012). The researchers found that experimental procedures did result in changes to implicit attitudes and explicit attitudes. However, the researchers found that the changes measured in implicit bias did not mediate changes in explicit attitude or behavior. In other words, though experimental procedures significantly impacted change to implicit attitudes, those changes did not appear to effect change in either explicit bias or behavior.

This finding provides important evidence bearing on the present discussion in two ways. First, unlike several other meta-analyses primarily investigating if and when implicit attitudes predict behavior, studying whether changes in implicit attitudes produce changes in behavior could potentially provide decisive evidence in favor of a causal relationship between these variables over and above a predictive relationship. Instead however, that is not what researchers found. Changes in the former did not produce changes in the latter. Though implicit states were predictive at levels found by past researchers, they did not mediate behavioral change, suggesting they association may not be causal. Second, a potential explanation for the finding that implicit attitudes are predictive of but do not produce changes in behavior is that discriminatory behavior could actually be explained by an underlying common cause of both implicit and explicit attitudes, without the need to posit a causal link with implicit attitudes to explain them. If this rather than implicit attitudes cause the discriminatory behavior in question, it further questions whether implicit attitudes themselves are causal and obviates the philosophical challenge to moral responsibility stemming from the supposed properties implicit states have.

Putting these pieces of positive and negative evidence together and evaluating this body of evidence as a whole, it appears that assuming causation is premature. Though there is continued disagreement between researchers, implicit attitudes likely predict a small proportion of variance such that only a small proportion of discriminatory behavior correlates with implicit attitude measures. The results are largely predictive and correlational in nature rather than causal and the strength of those associations uncovered to date casts doubt on the claim that implicit attitudes will be found to be significant causes of behavior. At present writing, then, the second premise of the ability argument is not adequately supported and should be rejected pending further evidence to the contrary.

At the same time, however, evidence in support of the claim that implicit attitudes are not causal also has limitations and should not be overstated. First, a limited number of studies, including those in Forscher et al. (2018)'s meta-analysis, directly compare implicit attitudes with behavioral measures. Further research may yet reveal robust causal links between implicit attitudes and other real-world behaviors that were not included. Second, behavioral change can be difficult to measure and this may be obscuring the relationship to implicit attitude change measured in the lab. Third, it is possible that small effects detected on the individual level could nonetheless still have an important and significant social impact in the aggregate (Greenwald et al. 2015) and that this is morally evaluable on the group level. For example, even if, as some researchers have suggested, IAT measures predict around 4% variance in measures of racial discrimination, this "represents potential for discriminatory impacts with very substantial societal significance" (Greenwald et al. 2015: 560). Thus, it is important to continue to study whether, and, if so, how implicit attitudes contribute to social outcomes. These caveats notwithstanding, however, research to date questions the likelihood that implicit attitudes will be shown to cause more than a small amount of behavioral variance on the individual level (see "Invalid Reasoning" for continued discussion).

Though there is currently insufficient evidence for the causal premise of the ability argument at present writing, future research may provide compelling evidence for it. To do this, research investigating moral responsibility for actions caused by implicit bias can be improved in several ways. First, researchers could identify the actions in question implicated in the argument against moral responsibility. The causal premise is largely idle until the specific actions implicit attitudes are said to cause are identified. Perhaps in future research this may result in subsequent challenges from implicit bias to moral responsibility that focus on specific discriminatory behaviors in a narrower subset of domains. Second, researchers could characterize with greater precision the nature and strength of the causal link between implicit attitudes and the behaviors in question to better understand the ethical significance of this link. This characterization will need to include a philosophical analysis of the strength of a causal contribution necessary to question control and moral responsibility. Third, researchers could continue to rule out plausible alternative explanations of correlations between implicit bias and behavior, for example, contextual factors, perceived control over a behavior, intentions, social norms, explicit attitudes that are not fully measured, or an underlying cause of both explicit and implicit attitudes.

4 The ability condition on moral responsibility

The normative core of the ability argument is the premise that moral responsibility for a behavior requires the ability to control it. The idea that ability limits responsibility is a foundational principle of ethics, often glossed in the slogan that "ought implies can". According to this principle, the presence of a genuine moral responsibility entails that an agent has the ability to fulfill it (Cicero and Edmonds 1856; Copp 2008; Feldman 1986; Hare 1965; Kant 1998; Van Fraassen 1973;

Vranas 2007). If an agent becomes unable to act, this rules out responsibility for the behavior. In other words, agents can never have a responsibility to do something after they become unable to do it. Unlike the first two empirical premises of the ability argument against moral responsibility, which can be confirmed or falsified by appealing to scientific evidence about how implicit attitudes work, the ability condition is a normative premise that requires philosophical arguments about the limits of responsibility and what morality requires.

One of the main arguments supporting the ability condition on moral responsibility is that ought implies can is an intuitively correct feature of moral psychology (Moore 1922; O'Neill 2004; Stocker 1971). However, there is good evidence that the ability condition on moral responsibility is not intuitively correct and that there are intuitive exceptions. Several philosophers have offered counterexamples demonstrating that it is sometimes intuitively correct to ascribe a moral responsibility to agents who lacks the ability to fulfill them (Graham 2011; Ryan 2003; Sinnott-Armstrong 1984; Spencer 2013). Perhaps the most well-known set of counterexamples against the condition from Walter Sinnott-Armstrong (1984) involve protagonists who limit their own ability to act in order to avoid fulfilling a promise:

> Suppose Adams promises at noon to meet Brown at 6:00 p.m. but then goes to a movie at 5:00 p.m. Adams knows that if he goes to the movie, he will not be able to meet Brown on time. But he goes anyway, simply because he wants to see the movie. The theater is 65 min from the meeting place, so by 5:00 it is too late for Adams to keep his promise (Sinnott-Armstrong 1984: 252).

According to the ability condition, the moral responsibility Adams has to fulfill his promise to Brown depends on having the ability to fulfill it by meeting Brown. Since he is not able to fulfill his promise after going to the movie, he no longer has a moral responsibility to fulfill his promise. Sinnott-Armstrong argues that Adams is morally required to fulfill his promise after he becomes unable to do so. Moreover, if obligations can be nullified simply by placing oneself in a situation that makes them impossible to fulfill, the ability condition risks trivializing the moral significance of genuine promises.

This intuition was recently confirmed in a series of experimental studies in moral psychology on ability and obligation judgments (Buckwalter 2017b; Buckwalter and Turri 2014, 2015; Chituc et al. 2016; Mizrahi 2015; Turri 2017). In one series of experiments, for example, researchers presented participants with materials similar to the case originally presented by Sinnott-Armstrong above (Chituc et al. 2016: Experiment 3):

> Brown is excited about a new movie that is playing at the cinema across town. He hasn't had a chance to see it, but the latest showing is at 6 o'clock that evening. Brown's friend, Adams, asks Brown to see the movie with him, and Brown promises to meet Adams there. It takes Brown fifteen minutes to drive to the cinema, park, purchase a ticket, and enter the movie. It would take 30 min if Brown decided to ride his bike. The cinema has a strict policy of not admitting anyone after the movie starts, and the movie always starts right on

> time. As Brown gets ready to leave at 5:45, he decides he really doesn't want to see the movie after all. He passes the time for five minutes, so that he will be unable to make it to the cinema on time. Because Brown decides to wait, Brown can't make it to the movie by 6 (Chituc et al. 2016: 23).

Participants agreed that "Brown ought to make it to the theater by 6" even though they strongly disagreed that "at 5:50, Brown can make it to the theater by 6". The researchers also found that participants strongly agreed that Brown should be blamed for failing to make it to the theater, and that these judgments about blame correlated with their obligation judgments. The researchers take this as evidence that obligations can sometimes intuitively extend beyond ability.

It might be objected that while protagonists like Adam and Brown do have moral responsibilities in the cases above, they did have at least some ability to fulfill them. For example, though Adams lacked the ability to fulfill his promise at the end of the story, he did have the ability to control this earlier in the story. Recalling the definition above, if it is reasonable to expect Adams to control his behavior at this earlier point, then Adams has "indirect" responsibility for his behavior at the later point. If this response to the counterexample is correct, it suggests that the ability condition is false because agents can be morally responsible for things beyond their direct control in virtue of being indirectly responsible for them. In other words, this suggests that moral responsibility is compatible with inability when the source of an inability can be traced back to the agent in certain respects, such as their free choices or negligence.

This response might be offered in defense of ought implies can, but is unlikely to support the argument against moral responsibility for implicitly biased behavior. It is likely that agents are morally responsible for implicitly biased behaviors in virtue of being at least indirectly responsible for them. For instance, agents could be responsible in virtue of failing to take prior courses of action to reduce biased behavior in the social sphere, including failing to educate themselves about the possibility of bias in environments that this knowledge is readily available (Washington and Kelly 2016). If agents are regularly responsible for biases they cannot directly control in virtue of being indirectly responsible for them, then the ability condition may be irrelevant to determining responsibilities in the majority of cases of implicitly biased behavior. In other words, defending ought implies can by disqualifying cases of morally evaluable inability as a counterexample to the inability rule will probably also end up proffering moral responsibility in the majority of bias cases. But, that said, suppose there are cases where no prior action, decision, or choice could prevent an agent from engaging in biased behavior. The question then arises, is the ability requirement intuitive even when agents lack both direct and indirect control of a behavior?

Further counterexamples to the ability condition suggest that the condition is not intuitively correct even when an inability to fulfill a responsibility cannot be traced to a prior ability. In the following case, proposed by Sharron Ryan (2003), for example, a protagonist is forced to act in a certain way due to an uncontrollable psychological compulsion:

> Suppose Sticky Fingers, your new friend, is keeping a little secret from you. She is a very serious kleptomaniac. She cannot ever help but steal the things she wants and she has an urgent desire for your CD music collection. After she comes to visit your house for the first time, you notice that all of your CDs are missing (Ryan 2003: 54).

According to Ryan, it is true that Sticky Fingers morally ought not steal the CDs, even though it is false she is able to stop from stealing them.[2] While kleptomania may serve as a genuine excuse for why Sticky Fingers deserves less blame or punishment than others for stealing, it does not rule out the presence of the moral responsibility not to steal. If an agent lacks the ability to refrain from stealing due to a psychological compulsion they cannot control or choose but still has a moral responsibility not to steal, this is another counterexample to the ability condition on moral responsibility.

Researchers have also replicated the intuition that moral responsibility can sometimes persist despite psychological compulsion and other psychological disorders among members of the general public (Buckwalter and Turri 2015; Turri 2017). In one study, for example, researchers presented participants with the following vignette where an agent is unable to act as a result of brain chemistry (Turri 2017: Experiment 4):

> A man is walking his dog in a public park. The dog is very violent. Given the current condition of the man's brain, it is impossible for him to warn people about the dog. As a matter of brain chemistry, it is literally impossible that he can warn people. He does not warn anyone. The dog bites someone (Turri 2017: 12).

Researchers found that participants strongly disagreed that the protagonist could have warned others or chose to warn others about the dog. The majority of participants also indicated there was a zero percent chance that the man would warn others. Despite these judgments, however, participants strongly agreed that the man had a moral responsibility to warn others. Researchers also found that participants ascribed blame to the protagonist ambivalently, or at about chance rates. These patterns of responses provide another counterexample to the ability condition and strongly question whether it is an intuitively true feature of moral responsibility.

It still might be objected that these protagonists had the ability to indirectly fulfill their obligations at some prior point. However, researchers have also replicated this finding in cases where an inability was lifelong, leaving no ambiguity about indirect ability. In one experiment, for example, researchers presented participants with the following case about an innocent bystander who is unable to act due to a physical disability (Buckwalter and Turri 2015: Experiment 5):

[2] Ironically, supporters of ought implies can are relatively unpersuaded by counterexamples involving psychological compulsion on the ground that they "are not true cases of incapacity, but, rather, are merely cases in which it would be very hard for the psychologically compelled person to refrain from doing what he or she is compelled to do" (Graham 2011: 342).

> Michael is relaxing in the park when he sees a small girl fall into a nearby pond. She is drowning and definitely will die unless someone quickly pulls her out. This part of the park is secluded and Michael is the only person around. But Michael's legs have been paralyzed since birth. As a result, Michael is not physically able to save the girl (Buckwalter and Turri 2015: 10).

Researchers found that the overwhelming majority of participants answered that Michael had an obligation to save the girl but was literally unable to do so. The researchers also found that participants strongly denied that Michael should be blamed for failing to fulfill this obligation. The researchers replicated this basic finding across many different conditions and cover stories. This result suggests that the ability condition is not intuitively true because having a moral responsibility is sometimes thought to be present despite the inability to fulfill it. This was shown even in cases where inability is lifelong and cannot plausibly be traced back to prior choices of the agent, as well as cases where agents are excused from blame.

While evidence suggests that the ability condition is not an intuitive requirement on moral responsibility, theorists have also argued that there are antecedent theoretical reasons to reject it that moved beyond what is intuitive (for a review, see Buckwalter 2017a; Talbot 2016). For example, one major theoretical reason, it is often suggested, to accept the condition is to preserve the conclusion that morality is motivational or action guiding. More specifically, the basic idea is that if morality requires us to do things beyond direct control, and impossible obligations cannot motivate or guide action, then morality would no longer be motivational or action guiding. Avoiding this result, it might be thought, provides a powerful motivation to accept the ability condition on moral responsibility.

This result does not follow for cases of implicit attitudes for three reasons. First, there is evidence that implicit attitudes are controllable and no evidence that the behavior they cause is impossible to control. Second, granting that both implicit attitudes are not directly controllable and that the impossible can never motivate, morality could still largely be motivational or action guiding. Even if some moral obligations fail to motivate (e.g. impossible ones), it does not follow that moral obligations fail to motivate. Third, it is false that impossible obligations fail to motivate. In the case of implicit biases, for example, recognizing that responsibilities persist in conditions where we lack direct control can be a powerful motivation to avoid future instances where we might lack it. For example, this might motivate us to adopt egalitarian policies or change the structure of our social institutions to avoid future discriminatory outcomes. The very fact that the ability condition is false may be what best motivates, as Ruth Barcan Marcus has called more generally, a "second-order regulative principle" that states that "as rational agents with some control of our lives and institutions, we ought to conduct our lives and arrange our institutions so as to minimize predicaments of moral conflict" (Marcus 1980: 121). These observations question a main theoretical reason to accept the ability condition, namely that rejecting it would render morality non-action guiding. It could be that rejecting the ability argument, in the long run, would make moral demands more motivational. There is reason to believe that this has already occurred, to some extent, as it may partially explain why millions of dollars have been spent on bias reduction training programs (Huet 2015).

To summarize, two distinct lines of research suggest that the third premise of the ability argument is false. One of the main arguments for the ability condition is that ought implies can is intuitively correct. However, several philosophers, cognitive scientists, and many ordinary speakers do not find the principle intuitive. Moral responsibilities are sometimes ascribed to agents beyond their direct or indirect ability to fulfill them. This finding was replicated across several narrative contexts, probing methods, types of moral requirements, and kinds of inability. The findings refute a central argument for the ability condition and provide some reason that it is false. Researchers have also questioned antecedent theoretical support for the principle beyond the claim that it is intuitive. One of the main theoretical reasons to accept the principle is that failing to do so would undermine the motivational nature of morality. However, rejecting the principle does not undermine this and may even partially explain why we are so motivated to avoid future discriminatory behavior through education, training, or institutional change.

Evidence against the ability condition also suggests one possible reason why it might have appeared intuitive to some. While researchers found that participants ascribed moral responsibilities beyond ability, they also discovered a nuanced connection between perceptions of ability and attributions of blame. For example, Chituc et al. (2016) found that agents were highly blameworthy when purposely limiting abilities. In the study presented above, Turri (2017) found that participants ascribed blame for unfulfilled responsibilities due to psychological compulsions ambivalently. Buckwalter and Turri (2015) found that blame was strongly denied to agents who were unable to act when this originated from a lifelong disability. This suggests that assessing the presence of a moral responsibility is different from evaluating agents for unfulfilled responsibilities, through assigning or excusing blame. Thus, it could be that the ability condition may have appeared intuitive because theorists sensed it would be wrong to blame protagonists in some circumstances. In other words, it could be that blame is more sensitive to facts about ability than judgments about responsibility are, and what appeared to be intuitive support for the ability condition was actually tracking this more nuanced connection between ability and individual cases of blame.

Lastly, distinguishing between moral responsibility on the one hand, and blame, punishment, or other social sanctions, on the other, may also lead to progress by improving our public dialog about implicit bias. Often public discussions of implicit bias are stymied by the perception that the identification of implicit attitudes is tantamount to an accusation of blame (Saletan 2016). Distinguishing between concepts of moral responsibility and blame may help correct this misrepresentation. The issue of whether a moral responsibility for behavior is present is separable from the issue of evaluating agents when those responsibilities are unfulfilled. According to this conceptual distinction, then, it is possible to acknowledge a responsibility for implicitly caused discriminatory behavior, while leaving open the question of whether different agents should or should not be blamed or otherwise sanctioned in particular circumstances in which ability is absent or impaired.

5 Invalid reasoning

For the sake of argument, suppose that each premise of the ability argument is true: implicit attitudes are uncontrollable, these uncontrollable states cause behavior, and moral responsibility for a behavior requires control of that behavior. Accepting these premises does not guarantee the conclusion that we are not morally responsible for implicitly biased behavior. To see why, notice that the premise regarding control pertains to the state p, not the action ϕ. For controllability to follow for ϕ from p, an additional premise is needed linking the properties that a cause has to the properties of its effect. Call this the *control transfer principle:*

> (CTP) If p causes ϕ then the inability to control p entails the inability to control ϕ.

This premise would render the ability argument against moral responsibility for implicitly biased behavior valid. However intuitive the principle might seem initially, though, there are strong reasons to reject it. In what follows, I argue that it should be rejected on logical, empirical, and conceptual grounds.

First, the principle should be rejected on logical grounds familiar to discussions of free will and moral responsibility. The present principle involves whether lacking control of a state entails that something that follows as a consequence of it was uncontrollable. Put this way, the principle closely resembles arguments against free will, and in particular, Peter van Inwagen's consequence argument for incompatiblism between determinism and free will (Campbell 2017; van Inwagen 1983, 1989). Briefly stated, the consequence argument is that:

> If determinism is true, then our acts are the consequence of laws of nature and events in the remote past. But it's not up to us what went on before we were born, and neither is it up to us what the laws of nature are. Therefore, the consequences of these things (including our present acts) are not up to us (van Inwagen 1983: 56).

The present discussion about implicit attitudes does not assume that determinism is true of course, but it does have a similar structure to van Inwagen's argument. Much like the laws of nature and past events, it might be thought, behavior is also not freely chosen if it is a consequence of implicit mental states beyond our ability to choose or to control. This connects the inability to control implicit attitudes to the inability to control behavior that follows from them.

Essential to the consequence argument is van Inwagen's well-known "Beta Principle" (van Inwagen 1989: 404–405). Given that "Np" stands for "p and no one has, or ever had, any choice about whether p" it says that:

> (Beta) From Np and N$(p \supset q)$, deduce Nq.

In short, if you have no choice p is true, and you have no choice that p implies q is true, then you have no choice that q is true.

However, the Beta Principle, as stated above at least, has been challenged by several theorists. Specifically, theorists have offered counterexamples demonstrating that it is invalid (Huemer 2000; McKay and Johnson 1996; Widerker 1987). One counterexample by Michael Huemer, for instance, asks us to imagine a device that shoots "R-particles" into a basket if and only if you freely decide to activate it (2000: 532–533). However you cannot control which half of the basket the particles will hit once it is turned on: there is a 50/50 chance that the participles will land on the left side or the right side of the basket. Further imagine that you decide not to activate the device. Given this initial setup, now consider the following propositions given by Huemer:

(A) No R-particle lands in the left half of the basket.

(C) No R-particle lands in the basket.

Since you have no choice which side of the basket would be hit if you turned on the device, you have no choice that A is true. For the same reason, you cannot choose whether A ⊃ C is true. Negating A ⊃ C is logically equivalent to A & ∼ C, and thus to falsify would require you to guarantee that R-participles hit the right side of the basket. Nonetheless, you can choose whether C is true, by simply choosing to turn on the device, which in turn, falsifies Beta and undermines the consequence argument.

This research suggests that if CTP is understood in terms of choice, then it is false. Understood in this way, having no choice about holding an implicit bias would imply that one has no choice about what follows as a consequence of it, namely, that the behavior it causes is inevitable. The consequence argument was one formal attempt to establish the conclusion that future outcomes are inevitable given that they are a consequence of determinism and other prior states of affairs that were not chosen. Counterexamples from this line of research suggest that, even after assuming determinism is true, the inevitability of ϕ would not follow from having no choice whether *p*. This gives us a logical reason to reject the similar pattern of inference operationalized in the discussion of moral responsibility for behavior caused by implicit attitudes we supposedly did not choose.

Researchers continue to debate whether alternative formulations of Beta are valid or whether there are other counterexamples to the consequence argument (Bailey 2012; Blum 2003; Vihvelin 2011). One response, for example, is that counterexamples can be avoided if the principle is adapted using modal verbs or the concept of alterability (van Inwagen 2000: 9, 2015: 19). Instead of invoking the inevitability of *p*, the revised principle invokes the notion of a "humanly unalterable truth that *p*", such that there is nothing a human being is or is ever able to do that "might" or "possibly" alter *p*. Further research is required to chart the implications of adapted Beta principles for the application to discussions of implicit bias. From the outset, however, such an adaptation appears unlikely to succeed in the present context. The ability argument against moral responsibility is premised on the claim that we lack the ability to control behavior caused by implicit attitudes, not that there is nothing humans can do that might alter behavior they cause. There are many things one can do to alter implicitly biased behaviors. And even if one is not fully convinced by the efficacy of experimental manipulations for implicit bias change on offer to date applied to various contexts, current evidence indicates, at the very least, that there is something an agent can do that might or possibly alter them.

It might be objected however, that framing the discussion in terms of conditionalization and logical consequence does not accurately characterize what is intuitive about CTP and what is at stake in the ability argument. Instead, it might be thought, CTP is a principle about how causation works in the world between specific sorts of mental states and behaviors. Refocusing the discussion in this way, it might be thought that CTP should be replaced by a more specific principle concerning behavior and implicit attitudes in particular. Call this the *implicit control transfer principle*:

> (ICTP) If implicit attitude p causes ϕ then the inability to control p entails the inability to control ϕ.

In other words, perhaps we cannot control actions caused specifically by implicit attitudes that we cannot control. After refocusing the discussion in this way, however, ICTP should be rejected in light of what we know so far about implicit attitudes and how they likely influence behavior. As reviewed above, there is insufficient evident that implicit attitudes are uncontrollable or causal. Furthermore, current findings suggest that if implicit attitudes do cause behavior, their causal influence will most likely be small. Of course, the fact that an effect is small does not necessarily mean it is not significant or important. But it does suggest something about causation and control on the individual level. Namely, we should expect that if implicit attitudes do cause behavior, they will almost certainly only ever partially cause it, along with all the other mechanisms, situational factors, external influences, beliefs, desires, and judgments relevant to human decision-making. As a result of this, even if a behavior was "caused" by an implicit attitude, and controllability did transfer from causal relata in the relevant way in the case of implicit attitudes specifically as in ICTP, it is doubtful that their causal contribution would be shown to come close to undermining control individuals ultimately had over their behavior.

In other words, it is unlikely that the causal impact of implicit bias on behavior is inevitable or unalterable, that the behavior in question owes its moral character to an implicit attitude, or that implicitly biased agents lack so-called "responsibility-level control" of their behavior.[3] To render these claims plausible, it would need to be demonstrated that these states impact us in ways that compromise significant amounts of freedom or agency. To show this, future research in both cognitive science and philosophy is needed to demonstrate both the causal impact of specific biases and to provide a theoretical framework specifying the threshold at which what are very likely to be small causal forces selectively undermine moral responsibility.

Some researchers have noted the issue of causal impact or behavioral prediction but have not recognized its full significance. Levy, for example, in arguing against moral responsibility correctly notes that "there is ongoing debate about what

[3] Thanks to Joe Campbell for discussion on this point.

proportion of behavior is predicted by implicit attitudes" (2017: 6). He does not engage with this debate, however, on the grounds that it only involves the matter of "how often" implicit attitudes cause morally relevant behavior rather than "whether" they cause it. Yet it is odd to bracket a debate about whether implicit attitudes can predict behavior for the purposes of analyzing an imagined case in which a behavior supposedly owes the entirety of its moral character to an implicit attitude. It is odd because the fact that there is a debatable relationship between these things questions the very idea any such behavior determined solely in that way exists. Conversely, it is more probable that no behavior owes its entire character to implicit attitudes and that vague causal language has exaggerated their influence on the control we ultimately retain over our behavior.

Lastly, the principle should be rejected on conceptual grounds stemming from ordinary evaluations of behavior and uncontrollable states more broadly. In the mental realm, it's not only possible but quite common to attribute the ability to control actions even though those actions happen to have been caused by mental states beyond control. For example, suppose hearing an insult causes an agent to become angry and this anger causes them to make an inappropriate comment towards a coworker. Further suppose this anger is beyond their control, and is the sole cause of their behavior. But that state of anger entails little about whether or not they had the ability to refrain from making the inappropriate comment towards the coworker or whether we hold them morally responsible for making it. In fact, controlling the behavior in spite of having the uncontrollable state is often referred to as "doing the right thing" and we often do hold others responsible for not doing that. And while we probably wouldn't blame the agent for becoming angry when insulted, we probably would blame or otherwise criticize the agent for behaving wrongly towards others in light of it.

The same is true for actions caused by many different kinds of mental states including explicit beliefs. For example, consider agents who hold explicit prejudicial beliefs as the result of some combination of upbringing, lack of education, and a lifetime of indoctrination. Continue to grant for the sake of argument that they are unable to change these beliefs and that these beliefs end up being the sole cause of prejudicial behavior. Tracking these features of a belief might help us to understand where the belief came from or why it continues to persist. But the causal etiology tells us very little about their abilities to have behaved otherwise. And the fact such agents cannot control that belief does not mean these individuals have no responsibility for prejudicial or racist actions.

Similar counterexamples exist in attitude control cases involving addiction. Suppose an agent has an uncontrollable desire to drink. And on many occasions, that desire does happen to be the sole cause of their drinking. Try as they might, they cannot change the desire. In fact, the desire might have deep physiological or psychological roots impervious to reason or intention. This desire, of course might make it extremely difficult for the agent to avoid drinking in excess, and in many cases, that might happen. But having an uncontrollable desire does not itself rule out the ability to act in addiction-discordant ways. The fact that addiction-concordant behaviors are not guaranteed by an addiction underlies positive treatments and outcomes of addiction.

If there can be control for behaviors that follow as a consequence of a range of mental states and other phenomena, including uncontrollable beliefs, desires, emotions, or even some cases of addiction, then it is unclear why the same cannot be said for behavioral consequences of implicit attitudes. Moreover, we regularly ascribe moral responsibility for behaviors these mental states cause. Of course, appealing to uncontrollable states can sometimes serve as legitimate excuses that mitigate blame or punishment for a behavior, in for example, some cases of addiction. But this is not to say that the possibility of moral responsibility being present is ruled out. These facts suggest that controllable behavior can follow from uncontrollable states.

Though there is reason to reject the claim that actions caused by uncontrollable states in general, or implicit attitudes specifically, entails that those actions are uncontrollable, this argument also has limitations. Rejecting this claim does not entail that there is no implicitly caused behavior we cannot control. For example, future research might find that some specific kinds of behavior are beyond control. The present discussion focuses primarily on complex or relatively high-level discriminatory behaviors of ethical concern, such as hiring or policing. However, it is possible that implicit attitudes could cause low-level behaviors that are themselves independently at the threshold of conscious awareness very difficult or impossible to fully control, such as minor adjustments to eye contact or tone of voice. To assess this possibility, future research is needed to isolate a unique causal role of implicit attitudes in these matters, evaluate whether behaviors such as speaking in a slightly lower tone of voice to someone is beyond our ability to control, and argue that these behaviors are morally evaluable with but not without that ability.[4]

To review, the ability argument against moral responsibility is invalid because it requires a premise transferring the controllability of attitudes to the controllability of behavior. Adding this premise renders the ability argument valid but attempts to defend it are not promising. As a logical matter, it remains unclear whether the fact that one proposition is uncontrollable entails another proposition is uncontrollable because it is a consequence of it. As an empirical matter, if implicit attitudes do cause behavior, evidence to date suggests they are likely to be just one among many other causes, including explicit attitudes, which questions their ability to undermine control. And lastly, the fact an uncontrollable state happens to cause a behavior does not rule out the possibility of control for that behavior because uncontrollable attitudes do not guarantee behavioral outcomes.

6 Conclusion

The question of ability and control looms large in scientific and philosophical research. According one picture of the mind in cognitive science, automaticity threatens to undermine the ability we have to control everyday decisions and

[4] Thanks to an anonymous reviewer for discussion on this point.

actions. According to one picture of moral responsibility in ethics, automaticity threatens the idea that we can be morally responsible for how we live our lives. The question of moral responsibility for implicitly biased behavior forces these two research traditions to a head. According to the ability argument against moral responsibility, if implicit attitudes that we cannot control explain a significant amount of discriminatory behaviors, and ability is necessary for moral responsibility, then we cannot be morally responsible for those discriminatory behaviors.

Fortunately, each premise of the ability argument against moral responsibility is unsupported and appears unlikely to be true. Evidence from cognitive science has not sufficiently demonstrated that we do not have control over implicit attitudes, that they cause behavior or that changing them leads to changes in our behavior. Neither is there sufficient support for the claim that behavior that is caused by an uncontrollable state is itself uncontrollable. Evidence to date may even suggest we actually do have the ability to control implicit biases and subsequently biased behavior. Pending future evidence, the rejection of these premises undermines the ability argument against moral responsibility.

The findings also question the priority of automatic processes over the conscious mind in human cognition and decision making. This fits a recent pattern of evidence across several domains of cognitive science suggesting that previously reported effects relating to the unconscious in various senses are either exaggerated or unreliable. For example, recent research suggests that contrary to conventional wisdom, procedural learning may take place with conscious awareness of implicit skills (Tran and Pashler 2017). Contrary to a widespread assumption of contemporary evaluation theory, automatic evaluations are probably not any less sensitive to validity information than deliberate evaluation are (Moran et al. 2017). Researchers have demonstrated that people can accurately predict and consciously report their own implicit attitudes (Hahn et al. 2014). Longitudinal studies find that individual differences in explicit measures are more stable over time than implicit measures are (Gawronski et al. 2017). Research casts doubt on social priming (Gerber et al. 2017; Pashler et al. 2012) or the reality of power posing (Jonas et al. 2017). And research questions the degree to which intelligence mindsets (Bahník and Vranka 2017) or stereotype threat (Finnigan and Corker 2016; Flore and Wicherts 2015) explain performance outcomes. Though these are but a few examples, and the issue requires a systematic review to properly address, the present research on implicit attitudes joins a trend that lends credence to view that it is the conscious rather than unconscious mind that often takes priority in determining behavior.

In philosophy and moral psychology, the role of ability might also be overstated. Theorists have questioned the notion that ability is always required for moral responsibility. Several philosophers have provided counterexamples in which an agent has a moral responsibility to do something they are unable to do. The findings suggest that while ability is clearly importantly related to responsibility and plays an important role in moral judgment and decision-making, positing it as a necessary condition is too strong and exaggerates its role in moral responsibility. These findings challenge the ability argument against moral responsibility because they suggest we can sometimes have a responsibility for things beyond our ability to control. At the same time, the research leaves open the question of whether and to what degree agents should or should not be blamed or punished for actions in different circumstances.

Finally, even if uncontrollable attitudes do cause behavior, the inference that the behavior they cause must be uncontrollable is an invalid one. For one, the inference relies on a control transfer principle heavily criticized in literature on free will and moral responsibility. For another, if implicit attitudes cause behavior they are likely to be but one cause among many that may make it more difficult but not impossible to act as morality requires. The use of causal language obscures this reality and renders ethical theories of implicit attitudes that exaggerate the causal relation idle. However, the fact that the behavioral contribution of implicit attitudes is likely small or that behavior could largely be explained by other factors does not make implicit attitudes any less important or ethically meaningful. To the contrary, attending to these facts clarifies why we are morally responsible for implicitly biased behavior.

Acknowledgements For helpful comments and feedback on previous drafts, I thank Jacob Berger, Carolyn Buckwalter, Joe Campbell, Shannon Dea, Mathieu Doucet, Daniel Kelly, Josh Knobe, Calvin Lai, and John Turri.

References

Ahmed, A. M., & Hammarstedt, M. (2008). Discrimination in the rental housing market: A field experiment on the internet. *Journal of Urban Economics, 64*(2), 362–372.

Allen, T. J., Sherman, J. W., Conrey, F. R., & Stroessner, S. J. (2009). Stereotype strength and attentional bias: Preference for confirming versus disconfirming information depends on processing capacity. *Journal of Experimental Social Psychology, 45*(5), 1081–1087.

Alston, W. P. (1988). The deontological conception of epistemic justification. *Philosophical Perspectives, 2,* 257–299.

Arpaly, N. (2003). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Bahník, Š., & Vranka, M. A. (2017). Growth mindset is not associated with scholastic aptitude in a large sample of university applicants. *Personality and Individual Differences, 117,* 139–143.

Bailey, A. M. (2012). Incompatibilism and the past. *Philosophy and Phenomenological Research, 85*(2), 351–376.

Bargh, J. A. (1999). The cognitive monster: The case against the controllability of automatic stereotype effects. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 361–382). New York: Guilford Press.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462–479.

Bargh, J. A., & Williams, E. L. (2006). The automaticity of social life. *Current Directions in Psychological Science, 15*(1), 1–4.

Bendick, M., Jackson, C. W., & Reinoso, V. A. (1994). Measuring employment discrimination through controlled experiments. *The Review of Black Political Economy, 23*(1), 25–48.

Bennett, J. (1990). Why is belief involuntary? *Analysis, 50*(2), 87–107.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246–263.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*(3), 242–261.

Blair, I. V., Ma, J. E., & Lenton, A. P. (2001). Imagining stereotypes away: The moderation of implicit stereotypes through mental imagery. *Journal of Personality and Social Psychology, 81*(5), 828–841.

Blincoe, S., & Harris, M. J. (2009). Prejudice reduction in white students: Comparing three conceptual approaches. *Journal of Diversity in Higher Education, 2*(4), 232–242.

Blum, A. (2000). The Kantian versus Frankfurt. *Analysis, 60*(3), 287–288.

Blum, A. (2003). The core of the consequence argument. *Dialectica, 57*(4), 423–429.

Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology, 7*(4), 765–786.

Brownstein, M. (2017). 'Implicit bias'. In N. Z. Edward (Ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/. Accessed June 19.

Brownstein, M., & Saul, J. (2016). *Implicit bias and philosophy: Metaphysics and Epistemology; Moral Responsibility, Structural Injustice, and Ethics* (Vols. 1 and 2). Oxford: Oxford University Press.

Bryan, C. J., Walton, G. M., Rogers, T., & Dweck, C. S. (2011). Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences, 108*(31), 12653–12656.

Buckwalter, W. (2017a). *Theoretical refutation of "Ought Implies Can"*. Manchester: University of Manchester **(unpublished manuscript)**.

Buckwalter, W. (2017b). Ability, responsibility, and global justice. *Journal of Indian Council of Philosophical Research, 34*(3), 577–590.

Buckwalter, W., & Turri, J. (2014). Inability and obligation: Compelling counterexamples to "Ought Implies Can". In *Buffalo experimental philosophy conference*, Buffalo, NY.

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0136589.

Campbell, J. (2017). The consquence argument. In K. Timpe, M. Griffith, & N. Levy (Eds.), *The Routledge companion to free will* (pp. 151–165). New York: Routledge.

Carlsson, R., & Agerstrom, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian Journal of Psychology, 57*(4), 278–287.

Castelli, L., Zecchini, A., Deamicis, L., & Sherman, S. J. (2005). The impact of implicit prejudice about the elderly on the reaction to stereotype confirmation and disconfirmation. *Current Psychology, 24*(134), 134–146.

Chapman, M. V., et al. (2018). Making a difference in medical trainees' attitudes toward latino patients: A pilot study of an intervention to modify implicit and explicit attitudes. *Social Science and Medicine, 199,* 202–208.

Chituc, V., Henne, P., Sinnott-Armstrong, W., & De Brigard, F. (2016). Blame, not ability, impacts moral "Ought" judgments for impossible actions: Toward an empirical refutation of "Ought" implies "Can". *Cognition, 150,* 20–25.

Cicero, M. T., & Edmonds, C. R. (1856). *Cicero's three books of offices, or moral duties: Also his cato major, an essay on old age, Lælius, an essay on friendship, paradoxes, Scipio's dream, and letter to quintus on the duties of a magistrate*. London: H.G. Bohn.

Cleeremans, A., & Jimenez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical* (pp. 1–40). Hove: Psychology Press.

Cooley, E., & Payne, B. K. (2017). Using groups to measure intergroup prejudice. *Personality and Social Psychology Bulletin, 43*(1), 46–59.

Copp, D. (2008). 'Ought' implies 'Can' and the derivation of the principle of alternate possibilities. *Analysis, 68*(297), 67–75.

Correll, J., Hudson, S. M., Guillermo, S., & Ma, D. S. (2014). The police officer's dilemma: A decade of research on racial bias in the decision to shoot. *Social and Personality Psychology Compass, 8*(5), 201–213.

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*(6), 1314–1329.

Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology, 92*(6), 1006–1023.

Dasgupta, N., & Asgari, S. (2004). Seeing is believing: Exposure to counterstereotypic women leaders and its effect on the malleability of automatic gender stereotyping. *Journal of Experimental Social Psychology, 40*(5), 642–658.

Dasgupta, N., & Greenwald, A. G. (2001). On the malleability of automatic attitudes: Combating automatic prejudice with images of admired and disliked individuals. *Journal of Personality and Social Psychology, 81*(5), 800–814.

Dasgupta, N., & Rivera, L. M. (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition, 26*(1), 112–123.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459.

Feldman, F. (1986). *Doing the best we can: An essay in informal deontic logic* (pp. 264–267). Dordrecht: D. Reidel Publishing Company.

Finnigan, K. M., & Corker, K. S. (2016). Do performance avoidance goals moderate the effect of different types of stereotype threat on women's math performance? *Journal of Research in Personality, 63,* 36–43.

Fischer, J. M. (2003). 'Ought-Implies-Can', causal determinism and moral responsibility. *Analysis, 63*(279), 244–250.

Flore, P. C., & Wicherts, J. M. (2015). Does stereotype threat influence performance of girls in stereotyped domains? A meta-analysis. *Journal of School Psychology, 53*(1), 25–44.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Michelle, H., Devine, P. G., et al. (2018). A meta-analysis of procedures to change implicit measures. *Open Science Framework*. https://doi.org/10.31234/osf.io/dv8tu.

Frankish, K. (2016). Who's responsible for this? Playing Double: Implicit Bias, Dual Levels, and Self-Control. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy: Metaphysics and epistemology* (Vol. 1, pp. 23–46). Oxford: Oxford University Press.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford: Oxford University Press.

Gapinski, K. D., Schwartz, M. B., & Brownell, K. D. (2006). Can television change anti-fat attitudes and behavior? *Journal of Applied Biobehavioral Research, 11*(1), 1–28.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures. *Personality and Social Psychology Bulletin, 43*(3), 300–312.

Gerber, A. S., Huber, G. A., & Fang, A. H. (2017). Do subtle linguistic interventions priming a social identity as a voter have outsized effects on voter turnout? Evidence from a new replication experiment. *Political Philosophy, 39*(4), 925–938.

Glaser, J., & Knowles, E. D. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology, 44*(1), 164–172.

Gonsalkorale, K., Allen, T. J., Sherman, J. W., & Klauer, K. C. (2010). Mechanisms of group membership and exemplar exposure effects on implicit attitudes. *Social Psychology, 41*(3), 158–168.

Graham, P. A. (2011). Ought' and ability. *Philosophical Review, 120*(3), 337–382.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the implicit association test can have societally large effects. *Journal of Personality and Social Psychology, 108*(4), 553–561.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the implicit association test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17–41.

Hahn, A., Judd, C. M., Hirsh, H. K., & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General, 143*(3), 1369–1392.

Hare, R. M. (1965). *Freedom and reason*. Oxford: Oxford University Press.

Hehman, E., Flake, J. K., & Calanchini, J. (2018). Disproportionate use of lethal force in policing is associated with regional racial biases of residents. *Social Psychological and Personality Science, 9*(4), 393–401.

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin, 31*(10), 1369–1385.

Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy, 43*(3), 274–306.

Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. *Psychology and Marketing, 27*(10), 938–963.

Huemer, M. (2000). Van Inwagen's consequence argument. *Philosophical Review, 109*(4), 525–544.

Huet, E. (2015). *Rise of the bias busters: How unconscious bias became silicon valley's newest target*. https://www.forbes.com/sites/ellenhuet/2015/11/02/rise-of-the-bias-busters-how-unconscious-bias-became-silicon-valleys-newest-target/#754ffb9f19b5. Accessed 25 July 2017.

Jonas, K. J., Sassenberg, K., Scheepers, D., & Wyer, N. A. (2010). Salient Egalitarian norms moderate activation of out-group approach and avoidance. *Group Processes & Intergroup Relations, 13*(2), 151–165.

Jonas, K. J., et al. (2017). Power poses—Where do we stand? *Comprehensive Results in Social Psychology, 2*(1), 139–141.

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed.). New York: Farrar, Straus and Giroux.

Kant, I. (1998). *Religion within the boundaries of mere reason and other writings*. Cambridge: Cambridge University Press.

Kelly, D., & Roedder, E. (2008). Racial cognition and the ethics of implicit bias. *Philosophy Compass, 3*(3), 522–540.

King, M., & Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy, 9*(2), 200–228.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., Greenwald, A. G., & Banaji, M. R. (forthcoming). Relationship between the implicit association test and intergroup behavior: A meta-analysis. *AmericanPsychologist*.

Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass, 7*(5), 315–330.

Lai, C. K., et al. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General, 145*(8), 1001–1016.

Legault, L., Gutsell, J. N., & Inzlicht, M. (2011). Ironic effects of antiprejudice messages. *Psychological Science, 22*(12), 1472–1477.

Lenton, A. P., Bruder, M., & Sedikides, C. (2009). A meta-analysis on the malleability of automatic gender stereotypes. *Psychology of Women Quarterly, 33*(2), 183–196.

Levy, N. (2014). Consciousness, implicit attitudes and moral responsibility. *Noûs, 48*(1), 21–40.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs, 49*(4), 800–823.

Levy, N. (2017). Implicit bias and moral responsibility: Probing the data. *Philosophy and Phenomenological Research, 94*(1), 3–26.

Machery, E. (2016). De-freuding implicit attitudes: Implicit bias and philosophy. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy: Metaphysics and epistemology* (Vol. 1, pp. 104–129). Oxford: Oxford University Press.

Mandelbaum, E. (2016). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs, 50*(3), 629–658.

Mann, N. H., & Kawakami, K. (2012). The long, steep path to equality: Progressing on egalitarian goals. *Journal of Experimental Psychology: General, 141*(1), 187–197.

Marcus, R. B. (1980). Moral dilemmas and consistency. *Journal of Philosophy, 77*(3), 121–136.

McKay, T. J., & Johnson, D. (1996). A reconsideration of an argument against compatibilism. *Philosophical Topics, 24*(2), 113–122.

Mekawi, Y., & Bresin, K. (2015). 'Is the evidence from racial bias shooting task studies a smoking gun? Results from a meta-analysis. *Journal of Experimental Social Psychology, 61,* 120–130.

Mendoza, S. A., Gollwitzer, P. M., & Amodio, D. M. (2010). Reducing the expression of implicit stereotypes: Reflexive control through implementation intentions. *Personality and Social Psychology Bulletin, 36*(4), 512–523.

Mitchell, G. (2018). Jumping to conclusions: Advocacy and application of psychological research. In J. T. Crawford & L. Jussim (Eds.), *The politics of social psychology: Public law and legal theory research paper series* (pp. 139–155). Charlottesville: University of Virginia School of Law.

Mizrahi, M. (2015). Ought, can, and presupposition: An experimental study. *Methode: Analytic Perspectives, 4*(6), 232–243.

Moore, G. E. (1922). *The nature of moral philosophy: Philosophical papers*. Abingdon: Routledge.

Moran, T., Bar-Anan, Y., & Nosek, B. A. (2017). The effect of the validity of co-occurrence on automatic and deliberate evaluations. *European Journal of Social Psychology, 47,* 708–723.

Moskowitz, G., Skurnik, I., & Galinsky, A. (1999). The history of dual-process notions, and the future of preconscious control. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 12–36). New York: Guilford.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474–16479.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2007a). The implicit association test at age 7: A methodological and conceptual review. In J. A. Bargh (Ed.), *Automatic processes in social thinking and behavior* (pp. 265–292). New York: Psychology Press.

Nosek, B. A., et al. (2007b). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*(1), 36–88.

O'Brien, K. S., Hunter, J. A., & Banks, M. (2007). Implicit anti-fat bias in physical educators: Physical attributes, ideology and socialization. *International Journal of Obesity (London), 31*(2), 308–314.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*(4), 421–433.

O'Neill, O. (2004). Global justice: Whose obligations? In D. K. Chatterjee (Ed.), *The ethics of assistance: Morality and the distant need* (pp. 242–259). Cambridge: Cambridge University Press.

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), 171–192.

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), 562–571.

Park, J., Felix, K., & Lee, G. (2007). Implicit attitudes toward arab-muslims and the moderating effects of social information. *Basic and Applied Social Psychology, 29*(1), 35–45.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of social distance? Failure to replicate effects on social and food judgments. *PLoS ONE, 7*(8), e42510.

Payne, B. K. (2006). Weapon bias. *Current Directions in Psychological Science, 15*(6), 287–291.

Payne, B. K., Brown-Iannuzzi, J. L., & Loersch, C. (2016). Replicable effects of primes on human behavior. *Journal of Experimental Psychology: General, 145*(10), 1269–1279.

Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry, 28*(4), 233–248.

Plant, E. A., Peruche, B. M., & Butz, D. A. (2005). Eliminating automatic racial bias: Making race non-diagnostic for responses to criminal suspects. *Journal of Experimental Social Psychology, 41*(2), 141–156.

Ramasubramanian, S. (2011). The impact of stereotypical versus counterstereotypical media exemplars on racial attitudes, causal attributions, and support for affirmative action. *Communication Research, 38*(4), 497–516.

Rusch, N., Corrigan, P. W., Todd, A. R., & Bodenhausen, G. V. (2010). Implicit self-stigma in people with mental illness. *The Journal of Nervous and Mental Disease, 198*(2), 150–153.

Ryan, S. (2003). Doxastic compatibilism and the ethics of belief. *Philosophical Studies, 114*(1–2), 47–79.

Sabin, J., Nosek, B. A., Greenwald, A., & Rivara, F. P. (2009). Physicians, implicit and explicit attitudes about race by MD race, ethnicity, and gender. *Journal of Health Care for the Poor and Underserved, 20*(3), 896–913.

Saletan, W. (2016). *Implicit bias is real. Don't be so defensive*. http://www.slate.com/articles/news_and_politics/politics/2016/10/implicit_bias_is_real_don_t_be_so_defensive_mike_pence.html. Accessed Jan 2017.

Saul, J. (2012). Skepticism and implicit bias. *Disputatio, Lecture, 5*(37), 243–263.

Schimmac, U. (2017). Reexamining Cunningham, Preacher, and Banaji's multi-method model of racism measures. https://replicationindex.wordpress.com/2017/01/08/reexamining-cunningham-preacher-and-banajis-multi-method-model-of-racism-measures. Acessed 13 June 2017.

Schwitzgebel, E. (2010). Acting contrary to our professed beliefs or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly, 91*(4), 531–553.

Sim, J. J., Correll, J., & Sadler, M. S. (2013). Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in the decision to shoot. *Personality and Social Psychology Bulletin, 39*(3), 291–304.

Sinnott-Armstrong, W. (1984). 'Ought' conversationally implies 'can'. *Philosophical Review, 93*(2), 249–261.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.

Spencer, J. (2013). What time travelers cannot not do (but are responsible for anyway). *Philosophical Studies, 166*(1), 149–162.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811.

Stewart, B. D., & Payne, B. K. (2008). Bringing automatic stereotyping under control: Implementation intentions as efficient means of thought control. *Personality and Social Psychology Bulletin, 34*(10), 1332–1345.

Stocker, M. (1971). 'Ought' and 'Can'. *Australasian Journal of Philosophy, 49*(3), 303–316.

Talbot, B. (2016). The best argument for "Ought Implies Can" is a better argument against "Ought Implies Can". *Ergo, 3*(14), 377–402.

Teachman, B. A., Gapinski, K. D., Brownell, K. D., Rawlins, M., & Jeyaram, S. (2003). Demonstrations of implicit anti-fat bias: The impact of providing causal information and evoking empathy. *Health Psychology, 22*(1), 68–78.

Teachman, B. A., Wilson, J. G., & Komarovskaya, I. (2006). Implicit and explicit stigma of mental illness in diagnosed and healthy samples. *Journal of Social and Clinical Psychology, 25*(1), 75–95.

Tran, R., & Pashler, H. (2017). Learning to exploit a hidden predictor in skill acquisition: Tight linkage to conscious awareness. *PLoS ONE, 12*(6), e0179386.

Turner, R. N., & Crisp, R. J. (2010). Imagining intergroup contact reduces implicit prejudice. *British Journal of Social Psychology, 49*(Pt 1), 129–142.

Turri, J. (2017). Compatibilism and incompatibilism in social cognition. *Cognitive Science, 41*(S3), 403–424.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1–9.

Van Fraassen, B. C. (1973). Values and the heart's command. *Journal of Philosophy, 70*(1), 5–19.

van Inwagen, P. (1983). *An essay on free will*. Oxford: Clarendon Press.

van Inwagen, P. (1989). When is the will free? *Philosophical Perspectives, 3,* 399–422.

van Inwagen, P. (2000). Free will remains a mystery. *Philosophical Perspectives, 14,* 1–19.

van Inwagen, P. (2015). Some thoughts on an essay on free will. *The Harvard Review of Philosophy, 22,* 16–30.

Vezzali, L., Capozza, D., Giovannini, D., & Stathi, S. (2011). Improving implicit and explicit intergroup attitudes using imagined contact: An experimental intervention with elementary school children. *Group Processes & Intergroup Relations, 15*(2), 203–212.

Vihvelin, K. (2011). Arguments for incompatibilism. In N. Z. Edward (Ed.), *Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/fall2015/entries/incompatibilism-arguments/. Accessed 15 Jan 2017.

Vranas, P. B. M. (2007). I ought, therefore I can. *Philosophical Studies, 136*(2), 167–216.

Wallaert, M., Ward, A., & Mann, T. (2010). Explicit control of implicit responses: Simple directives can alter IAT performance. *Social Psychology, 41*(3), 152–157.

Walton, G. M. (2014). The new science of wise psychological interventions. *Current Directions in Psychological Science, 23*(1), 73–82.

Washington, N., & Kelly, D. (2016). Who's responsible for this? Implicit bias and the knowledge condition. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy: Moral responsibility, structural injustice, and ethics* (Vol. 2, pp. 11–36). Oxford: Oxford University Press.

Widerker, D. (1987). On an argument for incompatibilism. *Analysis, 47*(1), 37–41.

Widerker, D. (1991). Frankfurt on 'Ought Implies Can' and alternative possibilities. *Analysis, 51*(4), 222–224.

Williams, B. (1973). Deciding to believe. In B. Williams (Ed.), *Problems of the self* (pp. 136–151). Cambridge: Cambridge University Press.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101–126.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283–301.

Yoshida, E., Peach, J. M., Zanna, M. P., & Spencer, S. J. (2012). Not all automatic associations are created equal: How implicit normative evaluations are distinct from implicit attitudes and uniquely predict meaningful behavior. *Journal of Experimental Social Psychology, 48*(3), 694–706.