Implicit Assumptions:
A case study on the IAT controversy

Scriptie ter verkrijging van de graad "Master of arts" in de filosofie
Radboud Universiteit Nijmegen

September 10th, 2017

F.J.W. Oude Maatman, s4074378
20,640 words (excluding references and footnotes)
Supervised by prof. dr. Jan Bransen
Philosophy of Behavioural Science

Hierbij verklaar en verzeker ik, Freek Johannes Wilhelmus Oude Maatman, dat deze scriptie zelfstandig door mij is opgesteld, dat geen andere bronnen en hulpmiddelen zijn gebruikt dan die door mij zijn vermeld en dat de passages in het werk waarvan de woordelijke inhoud of betekenis uit andere werken – ook elektronische media – is genomen door bronvermelding als ontlening kenbaar gemaakt worden.

Plaats: Nijmegen datum: 10 september 2017

Table of Contents

Abstract

In recent years, the implicit association test (IAT) has come under increasing scrutiny regarding its predictive validity. This thesis discusses possible shortcomings in the scientific process surrounding the IAT controversy from logical, methodological and sociological perspectives. First, a discussion of the current state of the controversy is given, after which the three perspectives are used to introduce several critiques of the IAT controversy. Four causes are identified: 1.) the lack of a supporting model for the IAT, 2.) the unsupported abduction of the IAT's creators to the current interpretation of the IAT, 3.) the influence of the implicit social cognition research program and 4.) a blind spot of social psychologists for underlying mechanisms.

1. Introduction

Imagine you are a police officer in San Francisco. Several years ago you have been tested by psychologists as part of a nationwide program of the US police force. Afterwards, they informed you that you suffer from a strong implicit bias against black people. This means that, unconsciously, you will treat black people worse than white people, even when you are strongly opposed to racism and discrimination on an explicit level (i.e., consciously). You are shocked; you do not consider yourself a racist, nor do you espouse racist beliefs or act in a racist manner at all. Therefore, you are motivated to get rid of this. Luckily, the psychologists also tell you that your implicit bias can be improved through an intervention. You go through the intervention process in order to improve, and are told that you need to repeat this annually to retain the beneficial effects. After several years, when the psychologists come in for one of your scheduled intervention at the precinct, you are told that this supposed 'implicit bias' actually might not really affect your behavior at all. Also, the test used to 'diagnose' you has been determined to be unreliable, and there are multiple different explanations for your (suddenly unreliable) score, besides your supposed racism. The intervention is shut down, and you are left wondering what happened – and why this was funded in the first place.

Of course, the scenario above is fictional. You are most likely not a police officer, I do not know anything about your implicit attitudes concerning black people and I am not sure how often you need to redo a bias-reducing intervention to retain its effect. Moreover, the anti-implicit bias training for police officers within the United States is most likely still in place[1].

[1] Implicit bias interventions aimed at lowering ethnic discrimination are being applied in the US police force. See Abdollah (2016). This intervention is most likely based on the intervention of Devine, Forscher, Austin & Cox (2012).

However, as absurd as it may sound, the rest is all true; without the mentioned fictions, the above is a short summary of an ongoing debate in the field of social psychology, which started last year. The *implicit association test* (IAT)[2], a well-known and oft-used psychological test with associated concepts such as *implicit bias*[3] and *implicit attitude*[4], has come under serious scrutiny as supposedly robust results concerning racial preference turn out to be based on unreliable evidence. Consequently, doubts have been raised concerning the interpretation of the test as well as the concept of *implicit bias*, next to the magnitude of their supposed predictive value for behavior[5].

This might seem like a 'scientific hiccup', something to be expected within the scientific process: some theories will be false, and finding out they are false is a form of progress as well. The sudden doubt concerning the IAT's predictions about discrimination however becomes a serious problem when we take into account that it not only has led to real-world training programs as stated above, but that there are also *thousands of published research articles* that use, mention or deal with the test in combination with racism, let alone the theory or concepts behind it[6]. This is not simply a possible refutation of a theory, which only has impact within the field. Thousands of hours of research could be determined to be a waste, and both governments and corporations might have spent thousands on ineffective training.

Meanwhile, the IAT is not alone; similar issues have been popping up elsewhere in social psychology. For example, another high-profile 'culprit' of irreproducibility is ego depletion theory[7], which luckily did not have nationwide interventions based on it – only a self-help book written by the authors of the theory[8]. It is likely other theories and independent studies will follow, as in a 2015 replication study only 39 out of the 100 replicated psychological studies showed the same significant effects as the original[9]. Academic psychologists and news outlets have dubbed this lack of reproducibility the *replication crisis*, and the scientific field is laboring to find an answer to it.

---

[2] See Greenwald, McGhee & Schwartz (1998).
[3] Implicit bias describes the possession of attitudes towards people or stereotypes associated with people outside of your conscious awareness. See "Implicit Bias" (n.d.).
[4] Implicit attitudes are defined as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects." See Greenwald & Banaji (1995).
[5] See Singal (2017) for an accessible discussion. For a more academic discussion, see Teige-Mocigemba, Klauer & Sherman (2010), which is quite complete even though being relatively dated, and chapter 2 of this thesis.
[6] A Google Scholar search reveals that the introductory article of the IAT, Greenwald et al. (1998), had been cited a staggering 9,099 times at the 7th of September 2017. According to Web of Science on the 8th of September 2017, this article has been cited 3,880 times.
[7] See Ferguson (2016) for an accessible discussion. See Hagger et al. (2016) and Curate Science (n.d.) for the scientific background.
[8] See Baumeister & Tierney (2012).
[9] See Open Science Collaboration (2015). The reported number (i.e., 39%) is based on the amount of studies that were subjectively rated to be successfully replicated. When looking at significance, only 36% percent of the studies provided statistically significant results.

Multiple causes have already been identified, which are mostly based on the current 'toxic' research environment. Questionable research practices[10], pressure to publish and publication bias[11] are the most cited explanations. For example, Nosek, Spies and Motyl argued in 2012 that pressure to publish and publication bias lead to questionable research practices. These are practices aimed at achieving statistically significant, and therefore publishable, results, which are necessary to further one's career. This leads to an inflation of false-positive results in literature, which can contribute to replication problems. Next to sociological arguments like these, methodological and statistical studies have shown possible ways in which questionable research practices can be conducted, and, in all likelihood, *are* being conducted[12]. In the last two years, a debate has started concerning whether replication studies are a panacea to all of social psychology's ills: multiple authors argue that a failure to replicate does not mean that the original research is invalid per se[13], even though this should not stop replication efforts from being undertaken.

However, as of yet, there remains a lack of investigation into *problems with the scientific process* of social psychology, which could also be causes of the replication crisis. Are the interpretations of data logically valid? Is there a proper clarity of concepts? Are used inductions and abductions warranted? Are auxiliary assumptions clear and verified? With this thesis I try to fill this gap in the literature, by treating the IAT and its current issues with predicting racist behavior as a case study, and scrutinizing the surrounding conceptual and philosophical framework. More specifically, I will look at the current controversy surrounding the IAT in detail and attempt to identify philosophical and conceptual mistakes that have contributed to its lack of replicability and the controversy as a whole. I will do this by first describing the history of the IAT, after which I will argue that the IAT paradigm can be treated as a Lakatosian research program, followed by a summary of its current critiques from inside the field of scientific psychology. After that, I will 'diagnose' the scientific process underlying the IAT through the usage of three perspectives that shed light on its current controversial status. Following this, I conclude with a final 'diagnosis' of the IAT.

---

[10] Research practices aimed at creating statistically significant results without committing data fraud. See John, Loewenstein & Prelec (2012), Simmons, Nelson & Simonsohn (2012) and Nosek, Spies & Motyl (2012).
[11] Pressure to publish refers to the academic pressure to publish research articles in order to progress (or even keep your job) as a scientist. Publication bias refers to the bias of journals towards novel, significant findings over replications or null findings. See Nosek et al. (2012).
[12] See Bakker, van Dijk & Wicherts (2012) for example, but also Simmons et al. (2012) and Ioannidis (2005).
[13] See Stroebe & Strack (2014), Cesario (2014) and Earp & Trafimow (2015). Notably, the latter base their argument against hasty falsification on old critiques against Popper's dogmatic falsificationism, instead pointing at a methodological falsificationism as proposed by Lakatos (1970) as a more correct framework for interpreting falsification through replication.

1.1 Disclaimer

While the goal of this thesis is providing a diagnosis of the scientific process regarding the IAT and its related concepts, I do not want to insinuate that the creators of the IAT or researchers who made use of the IAT for this purpose made grave mistakes, nor is it my aim to condemn them as 'incompetent researchers'. Instead, the goal of this thesis is to explicitly point out important steps in the research process which are often overlooked in social psychological research, using the IAT as a case study.

Similarly, I do not wish to suggest that racism is non-existent, nor that there can be no such things as 'unconscious racism', 'implicit bias' or 'implicit attitudes'. Instead, I wish to point out that the IAT is unlikely to measure any such thing due to critical mistakes concerning its model, aside from its supposed lack of predictive power. Throughout this thesis I refer to the usage of the IAT to measure prejudice and/or racism as a continuing example for several reasons: 1.) to stress the impact of the IAT, as this is the use of the IAT that has generated most interest, 2.) because this is the area in which it has the best predictive power over explicit measures, according to meta-analyses by Anthony Greenwald and Frederick Oswald[14], and 3.) because the White-Black paradigm is one of the most used IAT setups[15], and therefore also the most-discussed.

Besides these issues of interpretation, a knowledgeable reader could point out that my discussion and criticism of the IAT is incomplete, for instance due to missing several key alternative explanations of the IAT or clear empirical proof of its effects. In defense of this, it must be said that I have made a selection of those articles which have remained relevant and largely uncontroversial to date. For example, the usage of deliberate slowing strategies to influence IAT results is not mentioned as a problem for the IAT, because it has been solved in 2010[16]. Similarly, I do not include several articles that show significant correlations between IAT scores and behavior due to their inclusion in the meta-analyses used[17], or due to their rebuttal[18].

Lastly, while several of the arguments proposed in this thesis may be extended to other indirect measurements, the aim of this thesis is primarily to discuss the original IAT as published by Greenwald, McGhee and Schwarz in 1998, and covered in the various 'Interpreting and Using

[14] See Greenwald, Poehlman, Uhlmann & Banaji (2009) and Oswald, Mitchell, Blanton, Jaccard & Tetlock (2013).
[15] See Greenwald et al. (2009).
[16] See Cvencek, Greenwald, Brown, Gray & Snowden (2010).
[17] Idem footnote 15, but also Carlsson & Agerström (2016). These cover large amounts of ground.
[18] For example, McConnell & Leibold (2001) which has been refuted by Blanton, Jaccard, Klick, Mellers, Mitchell & Tetlock (2009), but remained an important basis for claims about IAT predictive validity until the refutation.

the IAT'-articles published by Greenwald and Banaji in the 2000s[19]. Variants on the IAT may be immune to critiques proposed in this thesis, for instance when a different scoring paradigm is used, or the test-procedure does not involve verbal associations.

[19] See Greenwald, Nosek & Banaji (2003), Nosek, Greenwald & Banaji (2005), Lane, Nosek, Banaji & Greenwald (2007), Nosek, Greenwald & Banaji (2007) and Greenwald et al. (2009).

2. The IAT: From Concepts to Controversy

Before I can discuss the IAT controversy as a philosophical case study, it is important to understand its history, key concepts and most important criticism. In this chapter, I will therefore describe the conception of the IAT and its related concepts, the most important of which are *implicit bias* and *implicit attitudes*. This is followed by an intermezzo, in which I introduce Lakatos' theory of research programs in order to introduce a sociological perspective which will be discussed further in later chapters. After that, an overview of the 'craze' and the controversies surrounding the IAT is given.

2.1 The IAT: Conceptual premises

Three years before the IAT was conceived, one of its creators and future proponents, Anthony Greenwald, professor at the University of Washington, published an article in collaboration with Mahzarin Banaji, professor at Harvard University and another future proponent of the IAT. The contents of this 1995 article would define the rest of their research careers, as it introduced the general notion of *implicit social cognition* and *implicit attitude* into social psychology[20], as an extension and integration of older psychological theories and new empirical findings[21]. *Implicit social cognition* was defined in this article as *'social cognitive processes that are inaccessible by introspection, that are caused by past experience of any possible type and which mediate current social behavior'*, and was introduced as a *'broad theoretical category that integrates and reinterprets established research findings, guides searches for new empirical phenomena, prompts attention to presently underdeveloped research methods, and suggests applications in various practical setting*s'. Greenwald and Banaji contrasted this *implicit social cognition* with self-reportable and introspectable cognition, which they dubbed *'explicit'*. Therefore, implicit social cognition is a broad category to refer to all social cognition which is not introspectable or self-reportable, which boils down to all social cognition whose activity or function we are introspectively unaware of.

In this sense, implicit social cognition could be seen as a redefinition of a Freudian sub-consciousness[22]. There are remnants of our pasts embedded in our minds which influence our behavior, which we remain unaware of. While Freudian thinking is currently considered 'debunked' for its lack of predictive power, research concerning the influence of unconscious processes and memory on social behavior had become an important part of social psychology by

[20] Of course, interest in subconscious processing existed before this, but Greenwald & Banaji introduce the notion of an 'implicit X' in order to describe unconscious variants of normally conscious cognitive processes or states which affect behavior. See Greenwald & Banaji (1995), p.5.
[21] See Greenwald & Banaji (1995), pp. 4 - 6.
[22] See Machery (2016), pp. 109-110, for an interesting discussion of this analogy.

the mid-eighties and early nineties[23].

The aforementioned *implicit attitudes* were defined by Greenwald and Banaji as a specific form of implicit social cognition. More specifically, they are defined as *'introspectively unidentified (or inaccurately identified) traces of past experience that mediate favorable or unfavorable feeling, thought, or action toward social objects'*. More simply put: implicit attitudes are introspectively unidentifiable 'likings' or 'dislikings' of a certain social object, which have an effect on your behavior. For example, say that through (traces of your) past experience cats have become associated with danger. When you now are confronted with a cat, you might avoid it or feel anxious. Yet, as per the definition of *implicit attitude*, you are not consciously aware of this relationship between cats and your behavior, thoughts or feelings. You might even consciously believe the complete opposite, such as that you like cats a lot.

An important difference between *implicit* and *explicit* processes becomes clear here. Implicit processes can affect your behavior without your cognitive control being involved and without your awareness of these processes happening. Explicit processes, in contrast, are introspectable and amenable to your cognitive intervention. In turn, this makes it very hard to measure any type of implicit phenomenon; you cannot directly ask a research subject about them, and without knowledge of how the implicit phenomenon works in the brain you cannot directly measure it either. Instead, you will have to rely on an indirect measure, which measures the implicit phenomenon's effects on behavior, emotion or thoughts.

This measurement problem was apparent to Greenwald and Banaji as well. Even though they had defined *implicit social cognition*, they still lacked the techniques to measure it - especially in the case of *implicit attitudes*. Only after three more years this problem was solved.

2.2 The IAT: Conception

In 1998, Greenwald, McGhee and Schwartz invented a way to measure the hypothesized *implicit attitudes*; the Implicit Association Test (IAT). The core of the test was, interestingly, based on a simple thought experiment. If one had to take a test in which one has to press button 1 for female names and male faces, and button 2 for male names and female faces, this test would be more difficult than a test in which faces and names matched genders on each side[24].

Greenwald[25] explained this by referring to associations; there exist strong associations between male names and male faces, and strong associations between female names and female

[23] See Greenwald & Banaji (1995), pp. 5 - 6.
[24] See Greenwald et al. (1998), p.1.
[25]..., McGhee & Schwartz (1998). I only mention Greenwald from here on forth, since he is the main author as well as the most prominent figure involved with the IAT debate of these three.

faces. Due to these associations, it is more difficult to quickly perform a task in which these associations are inverted than one in which they are not, since one has to 'overcome' the automatic responses following from them. From this, Greenwald concluded that the association strength between the categories sharing a button determines the difficulty of the experiment (i.e., relatively more strength leads to a quicker reaction), and thereby the time it takes to react correctly. This thought experiment provided the basis of the IAT model, shown schematically below.

IMPLICIT ASSOCIATION TEST

Sequence	1	2	3	4	5
Task description	*Initial target-concept discrimination*	*Associated attribute discrimination*	*Initial combined task*	*Reversed target-concept discrimination*	*Reversed combined task*
Task instructions	• BLACK WHITE •	• pleasant unpleasant •	• BLACK • pleasant WHITE • unpleasant •	BLACK • • WHITE	BLACK • • pleasant • WHITE unpleasant •
Sample stimuli	MEREDITH ○ ○ LATONYA ○ SHAVONN HEATHER ○ ○ TASHIKA KATIE ○ BETSY ○ ○ EBONY	○ lucky ○ honor poison ○ grief ○ ○ gift disaster ○ ○ happy hatred ○	○ JASMINE ○ pleasure PEGGY ○ evil ○ COLLEEN ○ ○ miracle ○ TEMEKA bomb ○	○ COURTNEY ○ STEPHANIE SHEREEN ○ ○ SUE-ELLEN TIA ○ SHARISE ○ ○ MEGAN NICHELLE ○	○ peace LATISHA ○ filth ○ ○ LAUREN ○ rainbow SHANISE ○ accident ○ ○ NANCY

Figure 1. Schematic description and illustration of the implicit association test (IAT). The IAT procedure of the present experiments involved a series of five discrimination tasks (numbered columns). A pair of target concepts and an attribute dimension are introduced in the first two steps. Categories for each of these discriminations are assigned to a left or right response, indicated by the black circles in the third row. These are combined in the third step and then recombined in the fifth step, after reversing response assignments (in the fourth step) for the target-concept discrimination. The illustration uses stimuli for the specific tasks for one of the task-order conditions of Experiment 3, with correct responses indicated as open circles.

Figure 1: A table depicting the setup of an implicit association test using the categories black names and white names, and an evaluation attribute, including original subscript. Copied from Greenwald et al. (1998).

In other words, the Implicit Association Test can be described as follows. The subject watches a screen, and is instructed to sort appearing stimuli from two categories by pressing one of two buttons; for example, pressing on the left button when a cat is shown, and on the right button when a dog is shown. After the first set of trials, in this case 'cat versus dog', a second dichotomy is sorted, which can be either another set of 2 categories (e.g., male faces and female faces) or an attribute (e.g., pleasant/ unpleasant, smart/dumb), using the same buttons and setup as before. In

our example, I will use the 'pleasant/unpleasant' attribute. The next set of trials becomes more complicated, as the subject has to do the two previous tasks *simultaneously*. Whenever an cat *or* a 'pleasant' word is shown, the left button is pressed, and when a dog *or* an 'unpleasant' word is shown, the right button is pressed. After this third, difficult set of trials, the first category is inverted: now you have to press the left button when a dog appears, and the right button when a cat appears on the screen. Then, the simultaneous task is presented once more, retaining the inverted first category.

For both instances of the simultaneous task reaction times are averaged, leading to a combined reaction speed of 'dog-unpleasant & cat-pleasant', and a second one for the inverse. The 'score' one has on the IAT is the difference between these two average reaction speeds. For example, say that the average response latency at the first task was 900 ms, while it was 800 ms at the second. In this case, the subject has a 100 ms difference - and it is this difference that is called 'the IAT-effect'[26], which is the primary output of the Implicit Association Test.

Summarized, this means that the IAT compares the *association strength* of two categories with a target attribute, such as pleasantness, or other categories, such as in the case of the 'names and faces'-example from the earlier thought experiment. In the example we used above, we can measure whether for a certain individual, dogs or cats are more strongly associated with pleasantness, by looking at the difference in reaction time between the two simultaneous tasks. If you are faster at the 'cat-pleasant and dog-unpleasant'-task, you supposedly have a stronger association between pleasantness and cats than between pleasantness and dogs, or a stronger association between unpleasantness and dogs than between unpleasantness and cats. Note that this is a *differential* association; it does not matter how fast or slow you are in both simultaneous tasks[27], only whether your responses are faster at one of the tests as compared to the other.

According to Greenwald, this technique made it possible to measure the *implicit attitudes* of individuals, the existence of which he and Banaji had hypothesized three years prior. He argues that a quicker average reaction time for the 'cat-pleasant and dog-unpleasant'-task, and thereby a stronger association between pleasantness and cats, indicates a relative difference in your unconscious 'liking' of cats and dogs. You *implicitly* like cats better than dogs, or in other words, you have a more positive *implicit attitude* towards cats than towards dogs. By itself, this might not seem like a very interesting finding. However, remember the difference between explicit and implicit; you might explicitly hate cats - yet, this test can tell you that you unconsciously like cats better.

Of course, using the IAT to measure unconscious cat/dog preferences is not the most

[26] Or D-score, as in 'Differential score'.
[27] Unless you vastly differ from the mean, for example by taking two tenths or ten seconds per answer.

pressing issue on any psychologist's agenda. Luckily, the IAT is a versatile test. In the same article in which he introduced the IAT, Greenwald also introduced its most famous and controversial use: measuring (implicit) black/white preference, or *implicit racial bias*.

2.3 The IAT: Implicit bias and the measurement of racism

Implicit racial bias, also simply known as *implicit bias* or *automatic racial preference*, is the name for a relatively positive or negative implicit attitude towards one ethnic group as compared to another ethnic group. It is measured like most other implicit attitudes; two different ethnic groups are used as categories, and are paired with an evaluation attribute (i.e., pleasant vs. unpleasant words). If you are faster at combining one group with the category 'pleasant' than the other, or faster at combining one group with the category 'unpleasant' than another, or perhaps even both; you possess an *implicit bias*, an implicit preference for one ethnic group over the other. An overly enthusiastic reader might, like the researchers, jump to a related conclusion: the IAT can measure (implicit) racism in subjects.

While I do not agree with this conclusion[28], one must admit this is at the very least an *intuitively plausible* step. First of all, a stronger conceptual association between, for example, black names and negative words than for white names and negative words, can easily be interpreted as a form of racism as it refers to a *preference* within the IAT framework. Preference of one race over the other can after all be argued to be close to the definition of racism[29]. Secondly, the IAT theoretically should be unaffected by social desirability[30], which makes it more methodologically suited for such a controversial subject than explicit questioning. Thirdly, there are precedents: previous research had reached similar conclusions with similar techniques. In 1983, Gaertner and McLaughlin[31] measured association strength similarly, yet instead of dividing the task over two different buttons, they measured reaction time for a yes-no question concerning whether the two words presented existed. In 1986, Gaertner and Dovidio[32] combined this with an evaluative measure and a yes-no question concerning whether the combination was 'always false' or 'could be true', again using reaction time as a measurement. They found that white subjects responded faster to positive traits after 'white people' primes than after 'black people' primes, and inversely

[28] See Chapter 3 for arguments supporting my view, which is not limited to the claim that the IAT is able to predict racist behavior.
[29] "Prejudice, discrimination, or antagonism directed against someone of a different race based on the belief that one's own race is superior", according to Oxford Living Dictionary. See "Racism" (n.d.).
[30] Responses to questions can be affected by cultural norms, such as those surrounding sexual activity, drug use and racism. People do not want to admit that they do not abide by the norms for fear of retaliation or ostracization, sometimes even when anonymity is guaranteed. Greenwald et al. (1998) argued this as well.
[31] See Gaertner & McLaughlin (1983), but also Greenwald & Banaji (1995).
[32] See Dovidio & Gaertner (1986).

with negative traits, which they interpreted as proof of aversive racism. In 1989, Devine[33] demonstrated similar effects of racism when priming subjects with African-American stereotypes and subsequently asking them to rate the hostility of a race-unspecified male; those primed with stereotypes considered the male more hostile. Greenwald knew of these experiments; they were mentioned in the 1995 article he co-wrote with Banaji, and were referenced in the 1998 article introducing the IAT as previous research that indicated the existence of unconscious racism.

Next to these older precedents, other researchers had already begun using the concept of implicit attitudes, and had started to prove their existence. For example, Dovidio et al. published a study in 1997, in which they argued that implicit attitudes against black people 'exist', once more using measurements based on reaction time like the IAT would a year later. Greenwald might not have known of this study, but Banaji did - she reviewed it, and had even offered advice[34]. In 1996, a study was published by Bassili, who argued that operative measures of attitude strength are more reliable than explicit measures due to them being less susceptible to 'extraneous influences' such as social desirability[35], and their ability to provide information about unconscious aspects of attitudes.

Given these arguments, the precedents and the later research, Greenwald arguably had enough reason to say that the IAT was suited for the measurement of unconscious racism. This was a major breakthrough; they had invented a tool that could measure a construct that had been nigh impossible to reliably measure before - and to top that off, it could also inform people about their unconscious position on one of the most controversial subjects of all time. Needless to say, Greenwald pounced on this opportunity, together with the aforementioned Banaji. This is evidenced by the 1998 press release[36] accompanying the IAT's introductory article, which proclaimed the importance of this construct in its first sentence:

> '*The pervasiveness of prejudice, affecting 90 to 95 percent of people, was demonstrated today in a Seattle press conference at the University of Washington by psychologists who developed a new tool that measures the unconscious roots of prejudice.* (...) *An important example is automatic race preference. A person may not be aware of automatic negative reactions to a racial group and may even regard such negative feelings as objectionable when expressed by others. Many people who regard themselves as nonprejudiced nevertheless possess these automatic negative feelings, according to Greenwald and Banaji.* (...) *While Banaji and Greenwald admitted being surprised and troubled by*

[33] See Devine (1989).
[34] See Dovidio, Kawakami, Johnson, Johnson & Howard (1997). In the acknowledgments, Mahzarin Banaji is mentioned by name.
[35] See Bassili (1996).
[36] See Schwarz (1998).

their own test results, they believe the test ultimately can have a positive effect despite its initial negative impact. The same test that reveals these roots of prejudice has the potential to let people learn more about and perhaps overcome these disturbing inclinations.'[37]

As can be seen in the quote, several large steps concerning what the IAT predicts were made here[38]. The IAT suddenly does not only measure a differential in associative strength between conceptual categories, it also predicts accompanying feelings and reactions: '*a person may not be aware of automatic negative reactions to a racial group*'. In one conceptual jump, we went from *'associations between verbal and/or visual categories'* to *'automatic negative reactions'*.

2.4 The IAT: Summary

Before we take a look at the controversy surrounding the IAT, I wish to shortly summarize the above. So far we have seen that the IAT is an extension of the theory of *implicit social cognition*, which is a broad theoretical category referring to all unconscious influences from memory on social behavior. It is aimed towards measuring the evaluative form of implicit social cognition; *implicit attitudes*. It does this by measuring reaction time differentials over different combinations of categories, such as 'cat-negative and dog-positive' and 'cat-positive and dog-negative'. Through this method it becomes possible to indirectly measure implicit attitudes, through measuring their effect on the reaction times. These implicit attitudes can have varied objects (i.e., what the attitude is about), and the most controversial variant is the *implicit racial attitude*, an implicit attitude towards ethnic groups.

Concluding, we see that *implicit social cognition* is the basis of the IAT methodology; it informed the search for implicit attitudes, and a way to measure these, which led to the creation of the IAT. However, note that *implicit social cognition* is not an empirical theory; it is a framework to guide research into implicit phenomena.

2.5 The IAT as part of a research program

Later in this thesis, I will introduce several perspectives on the IAT controversy, one of which is the claim that *implicit social cognition* as a theory has sociological implications for the IAT

[37] Quoted from Schwarz (1998).
[38] Next to that, it is noteworthy that it is not McGhee or Schwartz, one of the co-publishers of the IAT procedure, who take the stage with Greenwald, but that instead we see Banaji. Most likely this is explained by close involvement on Banaji's part with the creation of the IAT, as well as her earlier co-publication with Greenwald on implicit social cognition. This is partially evidenced by her being thanked in the article's acknowledgments for her comments.

controversy. In my view, these implications can be best described through the use of Lakatos' theory of research programs, through stating that implicit social cognition can be seen as such[39]. As it is useful to keenly remember the theory of *implicit social cognition* when this point is made, I wish to begin this argument here, and will continue and expand upon it in section 3.1.

Before I can argue that *implicit social cognition* is a research program, it is important to clarify what I mean by this term. In short, the concept of a research program refers to a sequence of theories characterized by a 'hard core' of shared assumptions[40]. This hard core is considered as above scrutiny, due to which falsifications of the theories within the research program instead are used to falsify - and then modify - the 'outer shell'. This 'outer shell' consists of auxiliary assumptions that have to be made in order to do research, such as assumptions about the accuracy of measurement instruments, but also of (ad hoc) assumptions to defend the core from too hasty falsification. The hard core of the research program namely has an implicit 'ceteris paribus clause' embedded in it[41]; X causes Y, *all other things being equal*. Consider for example a simple causation rule concerning gravity; 'all physical objects fall towards the center of the earth'[42]. However, in some cases physical objects might not do so; when a continuous force counteracts this, for example, or when an object is not under the influence of the earth's gravity well. In both of these cases, a new 'ad hoc' hypothesis could be introduced regarding this continuous force, effectively stating that the ceteris paribus clause has been broken (i.e., not all is the same), or that the measurement technique involved is faulty (i.e., the object actually *is* influenced by gravity). The ceteris paribus clause, together with the falsification of auxiliary assumptions, thereby work together. However, they also make the hard core unfalsifiable by itself - an ad hoc assumption can be generated each time to defend the hard core.

Lakatos has defended this use of ad-hoc explanations in science by referring to 'the positive heuristic'; as long as the ad hoc explanation leads to *novel* hypotheses, an ad-hoc defense of the hard core is legitimized. For example, it is possible to not falsify the concept of gravity when witnessing, for example, a helium balloon, as long as a new testable hypothesis is provided concerning the 'lack of gravity' working on the helium balloon. This hypothesis can then be tested - and when, like the previous experiment, the hypothesis is again falsified, another hypothesis can be generated. Yet, this process can only continue as long as new possible explanations can be generated. After which a research program is considered as 'degenerative'.

[39] Possibly for multiple reasons other than the similarity to Lakatosian research programs I perceive; implicit social cognition is not empirically tested, and effectively unfalsifiable as it has no practical implications by itself (i.e., ceteris paribus can be evoked). More about this will follow in the next chapter.
[40] See Musgrave & Pigden (2016).
[41] See Lakatos (1970), and Musgrave & Pigden (2016).
[42] I am aware of the fact that this is not a very accurate or up-to-date description of gravity, but it suffices to illustrate my point.

Unlike Popperian falsificationism, the inclusion of a positive heuristic prevents you from throwing out the proverbial baby with the bathwater: you do not risk falsifying the entire theory at stake all at once, and thereby avoid the possibility of losing the predictive value this theory *did have,* or *could* have had given the identification of additional laws.

It is time to return to the subject at hand. In the case of *implicit social cognition*, a hard core can be extrapolated from its definition: *'social cognitive processes that are inaccessible by introspection, that are caused by past experience of any possible type and which mediate current social behavior*'. From this follow the assumptions that a.) there exist cognitive processes that influence our social behavior, which we are introspectively unaware of, and b.) at least some of these implicit processes are influenced by our past experiences. In their 1995 article, Greenwald and Banaji support these two tenets of *implicit social cognition* with empirical evidence of several psychological phenomena that exhibit both introspective unavailability and causation by past experience, the most famous of which is priming[43]. Other support includes the lack of introspective access humans seem to have regarding their decision-making[44]. Together this support can be argued to make up the sequence of theories for *implicit social cognition*.

On this basis, Greenwald and Banaji then make a strong prediction: *'Individual differences in manifestations of implicit cognitive effects should be predicted by individual differences in the strength of theorized representations that underlie those effects'*[45]. This quote can be seen as the *main* prediction of the *implicit social cognition* research program. It proposes a causal relationship between implicit phenomena and behavior, which we already saw incorporated into the definitions of the previous sections. The IAT follows as a direct extension of this proposed causation; for lack of an ability to directly measure phenomena that are introspectively unavailable, it measures the 'manifestations' of implicit cognitive effects - in this case, the D-score, or reaction time differential - and extrapolates from these the strength of the underlying theorized representation, which are the supposed implicit attitudes. Within the *implicit social cognition* research program, this is like measuring the power used to kick a ball by measuring the speed of this ball as it hits a wall - an indirect way to gauge a causation, but a way to observe this causation nonetheless.

In the following sections and chapters, we will see whether this last belief holds up against scrutiny, and whether *implicit social cognition* can be fully treated as a research program. First we will, however, take a look at the 'splash' the IAT made in the world.

---

[43] Priming refers to the residual effect of a stimulus on the treatment of a following second stimulus. It is discussed further on p. 32.
[44] See Greenwald & Banaji (1995), pp. 5 - 7.
[45] See Greenwald & Banaji (1995), p. 6.

2.6 The IAT: Craze and its causes

Looking at the fact that the publication of the IAT procedure was accompanied by a press release and conference, it might be suggested that Greenwald and Banaji correctly predicted the enormous impact the IAT and the concept of *implicit bias* would have on the world. The IAT has not only made an enormous impact in the scientific and philosophical field[46], but it has generated ripples far beyond those of an ordinary psychological theory or measurement tool.

Especially within the United States, the test and its related concepts and predictions seem to have taken up permanent residence, mostly focusing on the White-Black preference application of the IAT. For example, training programs focused on the reduction of implicit racial bias have become part of government policy; not only for the American police force, but also for the American military[47]. Some American universities, like UCLA and Syracuse, conduct implicit bias trainings for their staff or provide implicit bias-related materials for self-study[48]. The IAT and implicit attitudes were mentioned during the first election of president Obama, as an explanation for relatively disappointing exit polls[49], and Hillary Clinton discussed implicit bias during one of the presidential debates with Donald Trump[50]. Even in the last four years, more than sixteen years after its publication, the race IAT is still brought up from time to time as a provocative headline[51], and popular media outlets all over the world have promoted the test[52].

How did a scientific tool transfer into the public debate at such a scale? In a critical 2006 analysis of the IAT, Fiedler, Messner and Bluemke argue that its popularity can be explained by its status as a *test*. As the IAT promises to measure (unconscious) prejudice, it not only is a valuable research tool but also fulfils a basic need: the need to reveal people's internal motives, desires and unconscious tendencies[53]. The IAT promises to reveal something about you which you are unaware of, but most likely have a strong opinion about; it promises 'a peek under the veil that your inept awareness cannot pierce, and shows you the truth', ugly as it may be. Even though the previous sentence is not exactly what is promised by the IAT, it is how many people perceive it, as can be evidenced by the media articles, researchers and even its creators treating it

[46] See footnote 6. For an overview of philosophical research on the IAT, I refer to Brownstein (2015) and "Reconsidering Implicit Bias" (2017).
[47] See the Picket (2017) and Abdollah (2016).
[48] See Weber (2016) and "Implicit Bias Resources" (n.d.).
[49] For example, see Rachlinski & Parks (2008).
[50] See the Washington Times (2016).
[51] As a small selection, e.g. Mooney (2014), Mooney & Viskontas (2014), Beres (2016).
[52] E.g., in the Netherlands we had the Volkskrant (2016) as a most recent example, but a quick Google search reveals mentions in Australia (Levy, 2012), England ("Are you prejudiced? Take the Implicit Association Test", The Guardian, 2009) and South Africa (Ngwetsheni, 2016), limiting myself to Anglophone countries.
[53] See Fiedler, Messner & Bluemke (2006), p. 78.

as such[54].

In a 2017 longread published by the New York Magazine, Jesse Singal argues that part of its success worldwide can be attributed to its *availability*. You can simply take the IAT online[55], and see for yourself whether you are unconsciously prejudiced or not. Later in his article, he goes a step further[56] by arguing that the story told by the IAT is so successful because it is *'politically palatable'*. According to him, the IAT tells us that implicit bias is a cause of many race-related issues, while also providing us with a means to detect it reliably. Therefore, the IAT seems to be a good method of tackling one of the main issues of our time: racism. Using the IAT in your research makes you part of the 'good side', 'the solution', as does acknowledging your own unconscious racism. Through research into the reduction of implicit attitudes and bias it might even lead to racism's possible extinction[57]. Or, at least, it *seemed* like it could do all these things.

2.7 The IAT: Prediction controversy

Given the success of the IAT, and the predictions made on its basis, one would assume that it is a very reliable measure with a proven connection to the concepts of implicit attitude and bias. Similarly, one would believe that implicit racial bias is proven to predict racist behavior. However, the IAT has been extensively criticized, or even proven *not to function as claimed,* on all the points that were just mentioned.

Before we discuss this, I first wish to point out that problems with the IAT are not caused by purposeful negligence of its creators and proponents. Anthony Greenwald, along with Brian Nosek and Mahzarin Banaji, has consistently published articles concerning the use and usability of the IAT[58] since its conception, even going so far to point out a 'Top 10' of things wrong with his own measurement instrument[59]. They are most certainly not closing their eyes for criticism either, given the many responses they have provided to critiques, and their willingness to solve, or agree with, identified problems[60]. Next to all that, until 2009 Greenwald frequently

[54] See any of the cited popular articles; e.g. Mooney (2014), Beres (2016). Also see Schwarz (1998) for proof of the indirect claims made by both the researchers as well as the writer of the article. Singal (2017) also cites many claims of both Greenwald and Banaji evidencing this, from personal correspondence, books and the literature.
[55] You can visit Project Implicit to take the test, a site which has been online since the IAT came out in 1998: https://implicit.harvard.edu/implicit/takeatest.html. The dataset it provided has been used in several articles by Greenwald and Banaji.
[56] See Singal (2017).
[57] See Schwarz (1998) as well for this suggestion.
[58] For example, see Greenwald, Nosek & Banaji (2003), Nosek, Greenwald & Banaji (2005), Lane, Nosek, Banaji & Greenwald (2007), Nosek, Greenwald & Banaji (2007) and Greenwald et al. (2009).
[59] Greenwald presented such lists in 2001 and 2004, one of which can be found online: https://faculty.washington.edu/agg/pdf/RevisedTop10.29Jan04.pdf
[60] For example, see their reply to Rothermund & Wentura (2004), Greenwald, Nosek, Banaji & Klauer (2005), or their reply to Oswald et al. (2013), Greenwald, Banaji & Nosek (2015).

updated a library on his personal website with articles concerning the various validity discussions of the IAT, facilitating debate by providing easy access to all critiques[61]. While commendable, this nevertheless has not yet solved several key problems with the IAT, even though we are nearing its 20th anniversary. In this and the following section, I will describe the key remaining problems.

The most well-known critiques of the IAT focus on problems of psychometrical importance, such as its lack of predictive validity[62]. Predictive validity is best explained as a measure of how well a test or measure predicts resultant behavior or other dependent variables. In the case of the IAT, this predictive validity varies greatly. Greenwald reports an average correlation between IAT scores and racist behavior of .236, while Oswald, using a more selective criterion for study inclusion, arrives at a correlation of .12, both of which are relatively low. Next to that, explicit measures (i.e., asking questions concerning racist attitudes) even seem to outperform, or perform equal to, the IAT when looking at correlations to race-related behaviors[63], and virtually all other areas of inquiry the IAT is used for, such as policy and consumer preference[64]. In fact, the only area in which the IAT outshines explicit measures is in MRI studies, where questions can easily be raised whether the observed activation spikes in the amygdala are indicators of a racist attitude, or emotional reactions of another kind. A 2006 meta-analysis by Carlsson and Agerström excluded doubtful discrimination measures like these. In their meta-analysis, they eliminate all discrimination measures that do not actually test for discrimination in their opinion (such as blinking responses and MRI studies), and find that there is no correlation between the IAT and the remaining discrimination measures overall. They then proceed to argue that the claim that the IAT can predict discriminatory outcomes has never actually been proven, due to methodological problems with the discrimination measures used and the lack of true experiments with the IAT[65].

As a final strong critique on the IAT's predictive validity, we can introduce yet another meta-analysis, Forscher et al. (2016)[66]. This meta-analysis aimed to assess the effectiveness of interventions aimed at changing implicit bias (i.e., IAT scores in the race IAT). Whilst their intention was to prove that implicit bias is malleable through training, an aim at which they succeeded, they however also found that a change in IAT scores doesn't lead to a significant change in racist behavior or the explicit bias .

---

[61] See Greenwald (n.d.).
[62] See Oswald et al. (2013) and Carlsson & Agerström (2016) for meta-analyses. There are several articles critically reinterpreting older publications concerning the IAT too however, such as Blanton, Jaccard, Klick, Mellers, Mitchell & Tetlock (2009). I am not mentioning these here as they are more strongly related to individual research than to the overall research program of the IAT.
[63] See Oswald et al. (2009), p. 183.
[64] See Greenwald et al. (2009).
[65] See Carlsson & Agerström (2016).
[66] See Forscher et al. (2016).

This leads to a startling conclusion; the IAT *has not been proven to predict racist behavior at all*, and if it does, it is at least *not better than the explicit measures which it is supposed to substitute.* Interestingly, the reaction of Greenwald, Nosek and Banaji to the meta-analysis by Oswald has been very calm[67]. They argue that even if the correlations of the IAT with outcomes are very low, this still can have significant effects on larger populations. This argument can be easily refuted however. First of all, correlations do not imply causation. That the two vary together (slightly) does not mean that implicit biases cause discriminatory behavior at all. Furthermore, if we translate the correlation coefficients into the, more regularly used and easier to interpret, effect size measurement r^2, we see that Greenwald is seriously mistaken. r^2 is also known as the coefficient of determination, and simply is the square of the correlation coefficient, which means that in the case of Greenwald's proposed correlation the r^2 is $.236^2 = .056$. The coefficient of determination measures the amount of variance in one of the variables that can be explained by the other in the sample; in this case, this can go both ways due to the unclear causation. This however does not mean that 5,6% of racist behavior which was included in Greenwald's meta-analysis can be explained using the score an individual had on the IAT, nor that 5,6% of IAT results can be explained by using the racist behavior of the individual. It means that for the individual, 5,6% of his racist behavior score or IAT score can be explained by using the other. This should be interpreted as a small nudge in the direction of discrimination at best[68] - if there even *is* a causation between implicit attitudes and discriminatory behavior to begin with[69]!

Another problem of the IAT is its test-retest reliability. In short, this refers to the correlation between the scores you get when taking the test twice, corrected for the length of time. If this is very low (i.e., the scores generally vary widely), questions can be raised concerning either the stability of implicit attitudes or the usability of the IAT for measuring them. In the case of the IAT the test-retest validity in general is determined to be approximately .55, with even worse numbers reported by Singal[70] and Gawronski, Morrison, Phills and Galdi[71]. This means that there is a relatively high chance that retaking the IAT will lead to a different result[72], allowing

---

[67] See Greenwald, Banaji & Nosek (2015). They have not yet reacted to Carlsson & Agerström (2016).
[68] If the other 94.4% are under conscious control, there is little chance that implicit biases will affect behaviour greatly.
[69] Meanwhile, Project Implicit - the online 'home' of the IAT - currently (June 1st, 2017) includes a disclaimer stating that no claim can be made surrounding the validity of the IAT's interpretations. See https://implicit.harvard.edu/implicit/takeatest.html
[70] Singal (2017) reports a number of .42, based on an assessment by Calvin Lai.
[71] See Gawronski, Morrison, Phills & Galdi (2017). They report a startling correlation of only .44 between two racial IAT's.
[72] If you are unfamiliar with correlations, I advise to look up a scatterplot with a correlation of .60 to see for yourself; https://allpsych.com/wp-content/uploads/2014/08/correlations.gif is a good example. Here you see in the bottom rows that several dots share the same x-coordinate ('first test') but not the y-coordinate ('second test'). Of course, this is not the most valid way of assessing the test-retest reliability of the IAT - it might be that you vary mostly between 'extremely heavily prejudiced' and 'heavily prejudiced', for example.

one to doubt whether the score delivered by the IAT is actually an *accurate* indication of implicit biases, or whether implicit biases are stable or not. All in all, the low test-retest reliability thereby reduces the importance one should give to an IAT outcome even further.

Nevertheless, this does not necessarily mean that the IAT itself is useless. Perhaps implicit attitudes are very unstable, causing both the low test-retest reliability and lack of predictive power. Maybe implicit attitudes don't affect behavior as strongly as was originally conceived, or IAT scores actually do not measure implicit attitudes. The latter two problems concern the theory and concepts *underlying* the IAT, and stretch further than the reasoning that was introduced in earlier sections.

2.8 The IAT: Methodological controversy

Through what mechanism(s) are implicit attitudes supposed to have an impact on behavior? How can we be sure that IAT scores are an accurate indication of implicit attitudes and of implicit attitudes only? These questions were part of the key problems pointed out in a psychometric and conceptual critique of the IAT, which was published in 2006 by Fiedler, Messner and Bluemke[73]. Their discussion of the IAT starts by mentioning the prevalence of implicit prejudice according to the IAT. This is incredibly high: 90 - 95% for anti-black prejudice amongst whites[74], for example. While this sounded alarming at the time[75], Fiedler[76] counters that it is perhaps the case that IAT scores indicating bias are a lot more common than actual racist (implicit) attitudes[77], meaning that the IAT is too sensitive as a measurement instrument. This leads to several conclusions, most important of which is that there might be causes for IAT scores indicating implicit attitudes *other than implicit attitudes themselves*[78]. This idea could partially explain the bad reliability and predictive power of the IAT mentioned in subsection 2.5, by introducing external moderating factors.

Fiedler continues his point with a theoretical critique. According to him, Greenwald and related researchers adhere to the idea that attitudes are evaluations (e.g. good, bad) associated

[73] See Fiedler, Messner & Bluemke (2006).
[74] See Schwarz (1998). Greenwald & Krieger (2006) published a more modest number of 64% of pro-white bias, but this was not corrected for race of the test-taker. Fiedler et al. (2006) use a number of 96% based on Greenwald et al. (1998) in their text.
[75] Note that the meta-analyses by Greenwald et al. (2009), Oswald et al. (2013), Forscher et al. (2016) and Carlsson & Agerström (2016) all were not published yet.
[76] I will use 'Fiedler' instead of 'Fiedler et al.' for textual reasons.
[77] See Fiedler et al. (2006), pp. 80 - 83.
[78] Fiedler et al. (2006) use several arguments to strengthen this claim; e.g. that other indirect measures correlate weakly with the IAT and that IAT scores are easily influenced by external factors.

with an object (e.g. white people)[79]. Measuring implicit attitudes then can indirectly be done by measuring the association strength between object and evaluation. In the case of implicit bias, a negative evaluation of a group then indicates a negative attitude towards it. Fiedler, however, rightly points out that mental associations and evaluations are a lot more complicated than this. One can for example have 'negative' associations such as associating the concept 'victimhood' with a certain group, causing one to behave as a protector towards that group, or one can simply have knowledge of stereotypes concerning that group whilst retaining a neutral attitude, yet still associating them with stereotypes[80]. Associations like those mentioned here also lead to an IAT result indicating bias[81]. These arguments show that 'negative' and 'positive' are not necessarily as clear cut in their effects on behavior as the creators of the IAT think them to be. Seeing 'evaluation' as a linear scale, in which all 'negative association' means that an unconscious racist attitude is present, is too simple.

Furthermore, many kinds of associations are possible, such as between the presented target stimuli and evaluation stimuli (e.g. 'George' and 'war'), between a target category and the abstract scale of 'evaluation' (e.g. 'white people' and 'negative'), between target stimuli and the abstract scale of 'evaluation' (e.g. 'Dick' and 'negative') or between a target category and evaluation stimuli (e.g. 'black people' and 'diamond'). Which combination of these is the IAT actually measuring? This is a large problem for the IAT. For example, do the evaluative stimuli in the IAT, like 'war', 'vomit' or 'diamond', map directly onto the evaluative categories they are supposed to represent (i.e. negative or positive), and *only* on these categories? It would be problematic if, instead of the category-evaluation association, one would also be influenced by the individual associations between the target stimuli and the evaluative words.

The question can also be raised whether the more abstract category-evaluation associations already *existed* in the test-taker, or have just been created ad-hoc for the task. For example, it is possible that association strengths rely on constant reinforcement. This might cause a faster reaction time in white people on the white-positive side of the task, due to self-referential effects, daily practice and cultural influences, such as advertising. However, there also would be a lack of strong associations between the other categories (i.e. white-negative, black-positive and black-negative), which would lead to a pro-white D-score.

The example above also shows that you only have to be faster (or slower) at *one* of the four sorting tasks to gain a bias-indicating result. This leads up to Fiedler's next argument,

---

[79] See Fiedler et al. (2006), p. 83. This can be confirmed when looking at the theoretical framework of *implicit social cognition* proposed by Greenwald, Banaji, Rudman, Farnham, Nosek & Mellot (2002), which proposes social knowledge structures based on linked concepts.
[80] See Andreychick & Gill (2012).
[81] See Uhlmann, Brescoll & Paluck (2006).

namely that the use of differential scores is ill-advised[82]. Non-attitudinal category associations[83] and other unwanted influences could have different effects on one side of the test (i.e. 'white-negative and black-positive' or 'black-negative and white-positive'), leading to a D-score that does not in any way resemble the actual implicit attitude. This also places a lot of weight on the chosen stimuli; if the evaluative stimuli chosen are more easily associated with one of the categories (e.g. 'gangster' and 'black people' or 'nazi' and 'white people'), or the target stimuli chosen are more easily associated with one side of the evaluation scale (e.g. 'Adolf' as a white name, or 'Barack' as a black name), this could bias results over all participants. After this claim, Fiedler shows that little to no attention is given to associations such as these; the focus lies on maximizing the evaluative strength of evaluation stimuli (i.e. as negative or positive as possible), whilst ignoring possible cross-category associations such as those mentioned above[84].

From these arguments, Fiedler concludes that the inferential interpretation of the IAT[85] is unwarranted; there are a lot of other possible interpretations which have not been refuted. Yet, then why do the IAT's creators believe in this interpretation? According to Fiedler, the creators of the IAT take for granted that attitudes can be inferred from reaction time latencies, *'by simply stating that an attitude results from every object-valence association and that the IAT taps into exactly this association*'[86]. The IAT therefore seems to be dependent on several assumptions. First of all, attitudes must result from single object-evaluation associations. Secondly, the IAT must measure this single association, and not any other association; in the case of the race IAT, this would be 'white people - evaluation' and 'black people - evaluation'. This means that the participant must make use of the category-evaluation associations only, leading to confounds when another cognitive strategy is used[87].

Fiedler then concludes that the IAT's link to implicit attitudes is only *assumed*, and that it will remain so until this link is proven in an experiment which could lead to its falsification. Interestingly, such an experiment has never taken place[88] as the IAT has only been used in correlational studies, and since there is no theoretical model that could form the basis for such an

[82] Fiedler et al. (2006), pp.93 - 98.
[83] Such as knowledge of stereotypes, familiarity, self-referential effects, etcetera.
[84] See Fiedler et al. (2006), pp. 89 - 92.
[85] I.e., the reaction speed differential is indicative of the implicit attitude.
[86] Quoted from Fiedler et al. (2006), p. 92. This can be proven by Greenwald et al. (2002), but also by Greenwald, Nosek, Banaji & Klauer (2005), p. 421; for instance quoting: *'Although Greenwald et al. (1998) used no theory of the structure of associative mental representations in presenting their interpretation of the IAT as a measure of association strengths...*'. Greenwald et al. (2005) then continues to argue that the IAT is theory-uncommitted.
[87] Several other strategies exist in the literature, and are shown to have different effects; Rothermund & Wentura's 2005 salience asymmetry interpretation, for example.
[88] Even after repeated pressure by Fiedler; see Friese & Fiedler (2010) and Fiedler & Hütter (2014) for example.

experiment. The only model related to the IAT[89] was published in 2002, but using this model would only complicate the IAT's results, as it proposes the possibility of split-concepts, which in short means that one could have a positive and negative concept of the same object at the same time. Which one of the two would be reached by the IAT (and whether another existed) would remain untestable. Yet, as Fiedler argues, a testable model leading to a full experiment, in which an experimental manipulation can be made, is necessary to be able to substantiate the claims made by the creators of the IAT regarding its ability to measure implicit attitudes.

Greenwald and Sriram disputed this necessity in 2010[90]. Firstly, they argue that such an experiment needs to manipulate association strengths (and thereby the implicit attitudes, according to their model), and then show that this change in association strength leads to different IAT results. However, at the moment it is not possible to measure association strengths directly, as we do not know how an implicit attitude is realized in the brain, nor do we have equipment sophisticated enough to measure minor changes in brain networks. This leads Greenwald and Sriram to conclude that such an experiment is unfeasible, because the results would be inconclusive; we can't be sure whether the association strength/implicit attitude would actually be manipulated. At the same time, they praise the value of correlational studies for validation, pointing out that these can be very strong when the causation is clear. Intelligence tests, for example, rely on correlation due to the inability to reliably manipulate intelligence in people, yet are considered to be very reliable.

Greenwald and Sriram's arguments do not hold up against scrutiny, however. First of all, placing the IAT on the same level as intelligence tests is unwarranted. There is no 'obviousness' in assuming that there is a link between reaction time measures and automatic negative associations, as there is with performance on an intelligent test and intelligence: such an 'obvious link' is exactly what is being disputed in the first place. Secondly, one could use other measures that 'tap into' association strengths, and use these for convergent validity - if such measures exist, of course. Designing an experimental approach for the IAT is maybe not possible *yet,* but it should be high on the priority list for people who wish to make claims such as 'IAT results predict racist behavior'. Thirdly, the defense offered by Greenwald and Sriram is a double-edged sword. If we cannot be sure that the IAT reliably measures the association strengths in the case

[89] See Greenwald et al. (2002). There exist alternative models that do not support the current IAT interpretation however, such as Mierke & Klauer (2001) and Rothermund & Wentura (2004). For a relatively complete overview, see Teige-Mocigemba, Klauer & Sherman (2010). Besides these, Brownstein (2015) relates several other psychological models to the IAT, such as the Reflective-Impulsive Model (RIM) and Associative-Propositional Evaluation (APE). As I have not been able to find an article in which these models are accepted by the creators of the IAT, nor general mention of them in relation to it, I have refrained from including them here. Attention must be drawn however to Amodio & Ratner (2011), who effectively have done what I will argue for in the rest of this thesis; looking at the neurological underpinnings of the IAT.

[90] See Greenwald & Sriram (2010).

of an experimental manipulation, how can we be sure that it does measure implicit attitudes in a normal situation?

Lastly, while the current lack of a testable model is a serious problem, it must be noted that the creation of such a model could nullify all the rebuttals I just gave. At the moment, we cannot create measures that indirectly or directly test for association strengths *due to the lack of a model* concerning these very things. *If*, and only if, a model existed, we could make predictions concerning what associations should consist of, and how they should be physically realized or in what way they could have effects in the world. This would, at the very least, lead to the possibility of *other* measurements of association strengths as well, allowing for convergent validity. Due to the lack of a model, the causality between association strengths, implicit attitudes and racist behavior remains untestable and thereby unfalsifiable.

2.9 The IAT: Conclusion

In this chapter we followed the IAT and its related concepts from conception to controversy. The IAT does not seem to live up to the claims proposed during its first publication, nor to the high value that is adhered to it as a predictor of discriminatory behavior in the media. The meta-analyses conducted by Greenwald and Oswald[91] show us that the predictive power of the IAT is mediocre at best, and not all strong enough to warrant the claims on its basis. Going even further, IAT scores indicating implicit biases are not yet proven to predict racist behavior at all, if we believe Carlsson and Agerström[92], and Forscher et al.[93]. And, even if we ignore these methodological problems, we are still left with the critiques of Fiedler, Messner and Bluemke[94], which show us that it is unclear what the IAT even measures, and whether its 'diagnosis' of prejudice is warranted at all; we don't know whether the IAT measures implicit attitudes because we don't know which associations these implicit attitudes would consist of.

In conclusion; the IAT currently seems to be in a theoretical limbo. An effect can be observed, namely a reliable difference in mean reaction times between different ethnicities, but what the cause of this effect is, or what is actually being measured, remains unclear. Yet, we have also seen the huge impact the IAT and related concepts have had outside of academia, and the claims that were made at its publication. How could this discrepancy have happened? In the following chapter I will attempt to answer these questions.

[91] See Greenwald et al. (2009) and Oswald et al. (2013).
[92] See Carlsson & Agerström (2016).
[93] See Forscher et al. (2016).
[94] See Fiedler et al. (2006).

3. The IAT: Three perspectives

In this chapter I introduce three lines of argumentation which will later be used to 'diagnose' the scientific process underlying the IAT. I will do so by approaching the IAT and its controversy from three perspectives, each focusing on a different aspect of this case study and highlighting a specific problem. I start with a sociological perspective, by continuing my introduction of the IAT and *implicit social cognition* into the Lakatosian framework of research programs. This is followed by a logical perspective, which focuses on abduction, and a methodological perspective zooming in on the hypothesized causal structure of the IAT. Then, I argue that the field of *social psychology* can be seen as a research program as well, in a continuation of the sociological perspective. Lastly, I continue the logical perspective by discussing the IAT in relation to *black box thinking* and inference.

3.1 A sociological perspective: I. the IAT as part of a research program, part 2

In section 2.5 I started my argument that the IAT is part of the *implicit social cognition* research program, whose hard core consists of the prediction that *implicit X'es* have an effect on social behavior, together with the assumptions that these implicit phenomena exist and are introspectively unavailable to us. Here I wish to conclude that argument.

In the previous chapter we have seen that over the last decade, the IAT has been under serious fire from two angles; its predictions and its conceptual and theoretical background. However, only the strongest criticism was included in this thesis. Other critiques included ways to fake your test results on the IAT, the influence of task orders, previous experience with the IAT, handedness, and so on[95]. All of these would be more or less parried by Greenwald, Banaji and other pro-IAT researchers such as Brian Nosek[96]. However, we have also seen that one important problem for the IAT – the lack of a testable model – has *never actually been addressed*. The 'implicit' assumption that *implicit attitudes* exist and directly affect behavior, and that the IAT can measure these implicit attitudes, thereby has avoided scrutiny altogether. Instead, the discussions and critiques mostly focused on problems *surrounding* this alleged measurement; faking strategies and possible outside influences on this measurement were discussed. As another example, a large debate concerned the question whether *culture* is measured instead of personal biases[97] – which

[95] See Greenwald (n.d.) for Greenwald's own log of criticism on the IAT, which holds all of these examples.
[96] Through published replies and refutations, such as Dasgupta, McGhee, Greenwald & Banaji (2000) and the aforementioned Greenwald & Sriram (2010).
[97] See Greenwald (n.d.).

could lead to the IAT still measuring a relevant implicit attitude affecting behavior, albeit cultural instead of personal. While this debate may seem critical, it actually only targets the *causes* of the supposed implicit attitude; is it a suppressed 'actual attitude' or created through culture?

The events described above can be seen as characteristic of a research program; the hard core is above scrutiny. Besides Fiedler, Messner and Bluemke[98], no one has truly questioned the IAT's core hypothesis that verbal associations can influence behavior. Next to that, instead of falsifying the hard core, auxiliary assumptions concerning the IAT were refuted or changed, for example by changing the scoring algorithms[99] and inventing a reliable way to identify the use of faking strategies[100]. This is an example of the 'positive heuristic' at work. Given these facts, I believe that treatment of *implicit social cognition* as a research program is warranted.

When considering *implicit social cognition* as a research program, we gain several insights into the IAT controversy and tools for its analysis. Firstly, it shows that we should not consider the IAT in isolation of the *implicit social cognition* research program, as it is its methodological offshoot. Not only the assumptions of the IAT itself are relevant in understanding the causes of its overestimation and subsequent controversy, but those of *implicit social cognition* as well.

Secondly, this perspective shows that the IAT does not 'stand alone'. If it fails, by extension *implicit social cognition* can be said to be under serious pressure. A possible 'failure' of the IAT however does not mean that the hard core of *implicit social cognition* is falsified. After all, as pointed out in section 2.7 and 2.8, it is possible that the IAT is very unreliable in its measurement of implicit attitudes, due to inclusion of other, confounding effects on reaction time. The belief that the IAT correctly measures implicit attitudes is an *auxiliary assumption* which can be scrapped to protect the hard core, and, perhaps, with enough ad hoc hypotheses, the IAT might be saved as well[101].

Thirdly, and most importantly, it allows us to reappraise the reasons for the IAT's overestimation, adding to the existing arguments described in section 2.3. Interpreting the IAT's results as indicating the root of prejudice can be seen as a symptom of the confidence in the underlying research program. A form of tunnel vision after all seems to have occurred here: the found IAT effect was immediately seen as proof for the research program – while many other hypotheses could have explained the same results. For example, instead of jumping the gun with terms such as 'prejudice' and 'negative implicit attitudes', it could have been argued that the negative verbal associations causing the difference in IAT-results might have been strongly related to knowledge of the history of black people (e.g., slavery, struggle for civil rights) instead

[98] See Fiedler et al. (2006).
[99] See Greenwald et al. (2003).
[100] See Cvencek et al. (2010).
[101] Creating a testable model would be highly useful for this aim.

of to prejudice towards them, as was mentioned in section 2.8. Next to that, it might be completely unrelated to affective differences towards groups, instead being caused by familiarity differences with the IAT's used stimuli, such as the names or facial structures. Another option would be that the negative affect involved is not in fact hostile, but rather guilt or shame, or that the measured verbal associations do not have an influence on behavior or thought outside of reaction times. Lastly, it is possible that an IAT score indicating prejudice is influenced by multiple factors at once: actual hidden prejudice, compassion, knowledge, familiarity and culture. In this last case, groups with prejudice-indicating scores are most likely to discriminate simply because the people who actually discriminate are more likely to end up in them, while the IAT result itself is not a good predictor of prejudice at all.

Some of these options were proposed in critiques of the IAT and some even by the researchers themselves in their introductory article. Yet, the jump to 'prejudice' and a causal relationship between implicit attitudes/verbal associations and discriminatory behavior was still made by the IAT's creators in the press release by Schwarz[102]. This can be argued to be the 'hard core' of the *implicit social cognition* research program at work; the (multi-interpretable) evidence is interpreted as supporting the hard core. However, instead of supporting it, the evidence only did not *falsify* the hard core; that took several decades, as we have seen in chapter 2. And as expected, instead of dropping the *implicit social cognition* research program, the IAT's proponents mended the outer shell, by for example changing the scoring algorithm underlying the IAT's results. Following that, they seem to have made use of the *positive heuristic*, postponing the pending falsification with alternative explanations for falsifying results, especially by criticizing the methodology of critiques.

Yet, one fact remains. In the face of a large amount of possible hypotheses, only one seems to have been chosen as 'valid' by the IAT's creators – even though they recognized the possibility of other hypotheses in the actual article introducing the IAT[103]. In the next section, we will change perspectives and zoom in on this irregularity.

3.2 A logical perspective: I. Abduction

The jump made by Greenwald and Banaji in 1998, as described above, can be considered as a mistake in *abduction,* also known as *inference to the best explanation*[104], or, colloquially, 'reverse inference'. Abduction is the inverse of deduction; instead of inducing a consequence from a cause

[102] See Schwarz (1998).
[103] Such as familiarity, which they tried to partially rule out. See Greenwald et al. (1998), p. 1477.
[104] See Douven (2017).

and a law, a (probable) cause is induced from a consequence. It is also known as 'post hoc ergo propter hoc'. An abduction has the following logical structure:

1. X is observed
2. If Y is true, then X would logically follow, where Y is the best explanation of X out of a set of possible explanations Z
3. Therefore, Y is (probably) true

While this is not a logically valid inference[105], this type of reasoning permeates our daily lives. For example, say you just woke up and find that the bag of bread on your kitchen top has a hole in it. Also, there's crumbs all around and a piece of bread has been gnawed on. You will conclude that there are mice in your house - but this is not a *necessary* logical conclusion. There might be nocturnal ants that eat through plastic and bread, or maybe a vermin eradicator has broken into your house and done this in order to gain more business. Other options are not excluded by your reasoning or the circumstances. However, these other possible explanations seem very unlikely in comparison, which is why most people will infer the existence of mice from such a situation; due to its likeliness, it is considered the *best explanation* available. Nevertheless, this does not mean that the other possible causes are not the case.

This process of abduction also lies at the core of *scientific reasoning*[106]. Given a certain phenomenon and our relevant background knowledge, we construct an interpretation concerning what is going on: a hypothesis. For example, when asking ourselves why things fall down and not up, we could say that they are seeking their natural place in the world, or that the gravity of the earth is working on the object. Dependent on our background knowledge, some interpretations seem more likely, or *better*, than others. When considering the example of the opened bread bag, we see that some explanations are plain silly, yet theoretically possible. However, there are also possible explanations that can immediately be refuted based on your background knowledge - that your chess set did it, for example.

Scientific reasoning follows a similar process; explanations are generated based on our background knowledge of previous experiments and confirmed phenomena. It is however also slightly stricter: scientists do not generate a random set of guesses and pick one. They generate a set of explanations based on previous research findings and theories about the phenomenon at

[105] The strength of the conclusion after all depends on the (impossible to determine) completeness of the set Z; you might abduce to the best explanation of a bad lot. Next to this, the conclusion depends on your explanation criteria.
[106] See Douven (2017), Addendum: Peirce on Abduction.

stake, and compare these[107]. A hypothesis therefore is not random; it is (hopefully) based on a solid amount of research, knowledge and comparison. Nevertheless, they are not derived truths; they remain a 'guess' which is yet to be confirmed, albeit a highly educated one. Luckily, a hypothesis does not have to remain in 'truth value limbo'; it can be confirmed, refuted or altered when new relevant information becomes available, such as new experimental outcomes or theories.

The above may have come across as elementary. Of course, a hypothesis is unconfirmed, and due to it being man-made it is dependent on the criteria and knowledge used by its creator. Yet, this points us to an interesting possibility: in some cases, several interpretations are, in fact, equally likely, yet we may only consider *one*. While this is not too problematic for daily life, for example by concluding the existence of mice in your house from supposed traces of mice, in science there is no such luxury. As we have after all seen, a conclusion rising from abduction is not guaranteed. In order to prove it, we have to 'check', for example by conducting experiments that can confirm the inferred causal relationship, or by excluding other possibilities. When we do not do this, yet simply continue with our research as if nothing happened, we are effectively building a 'cloud castle' - a theory with very shaky foundations.

This is the mistake made when the IAT was published, and which continued to permeate discussions of the IAT later. Instead of acknowledging that the IAT effect could have been explained from various other angles, Greenwald and Banaji jumped to the most readily available two conclusions[108], namely 'this is proof for implicit social cognition' and 'these results explain prejudice'. They refuted some of the competing hypotheses later[109], but they never tested their own explanation directly, as we saw in the previous chapter, instead relying on correlations between IAT results and measures of racist behavior. This explanation however followed the same abductive inference pointing towards *implicit bias*: 'People who score as prejudiced against blacks on the IAT have negative implicit attitudes/implicit bias, and therefore also discriminate against blacks'. Their *inference to the best explanation* therefore seems to have heavily favored what they *wanted to find*[110]. This viewpoint can be strengthened when we take into account that, as we saw in section 2.2, the IAT was *created in order to measure implicit attitudes*. The abduction towards *implicit attitudes* therefore can be argued to also have been influenced by the fact that the IAT was

[107] This is a difficult question in philosophy of science, see Douven (2017). Which criteria are used in order to compare possibilities is dependent on the field in question, and presumably preferences of the researchers involved - for example, a slight preference for explanatory power over parsimony. Another answer to the 'how' of hypothesis-forming could be generated from Bayesian epistemology.
[108] Most likely also due to the higher chance of publication when such a prediction was made, and by other factors already described in section 2.3.
[109] See Dasgupta et al. (2000), for example, which refutes the familiarity hypothesis.
[110] This could also be seen as a good example of confirmation bias.

meant to measure these in the first place, pushing the explanation and interpretation given even more to the foreground.

The most important role however must be given to the fact that the hypotheses generated in 1995 were confirmed; the IAT's first results lined up with the predictions made then, namely following the general lines of *implicit stereotypes, attitudes and bias*. This support was seen as *proof* for the IAT's effectiveness at tapping into *implicit attitudes*, without considering all the other aforementioned possible hypotheses that could have given similar results.

This points not only to a mistake in logic in the IAT, namely the continuation of abductive inference in spite of this being insufficient, but also to a mistake in the scientific process leading up to and following its creation. The abduction from IAT results to 'implicit attitudes' and 'social behavior' should have been questioned and proven *before* claims surrounding the IAT's effectiveness for predicting social behavior and implicit attitudes were made. Yet, the main published critique in this area, the article by Fiedler, Messner and Bluemke, appeared 8 years after the fact, and never was properly reacted to by IAT proponents[111].

The only way to make this abduction solid was mentioned above: showing that it is *better* than the other explanations, by directly relating it to the world. At this point, the IAT is only an *abstract* hypothesis, without empirical roots strangling the opposition. Showing that several alternatives are even weaker, is not sufficient to justify the claim that the IAT's current interpretation is valid. The claims need to be verified *on their own*.

In my view, the best method to achieve this would be forming an underlying model, which explains the steps from IAT results to verbal associations to implicit attitudes to discriminatory behavior, followed by the empirical testing of its predictions other than 'IAT results can predict racist behavior'. Without such a tested model, the data found by the IAT remains multi-interpretable, and the validity of the current interpretation remains clouded. Excluding other possibilities does after all not prove another possibility; there is an effectively unlimited amount of possible explanations for any type of phenomenon. All that therefore effectively has been found in IAT research so far, is that there is a reliable difference in reaction time differentials between people from different ethnicities when taking the race IAT.

[111] This will be expanded upon in section 3.4.

3.3 A methodological perspective: the IAT's hypothesized model

So far, the hypothesized model of the IAT has been mentioned and discussed several times, especially when proposing that a stronger model should be created in its place. Nevertheless, we have not scrutinized the current hypothesized model of the IAT itself. In this section I wish to describe this model, and then explicitly show why it is insufficient. After that, I introduce an alternative explanation for the found, reliable IAT results - namely the large prevalence of bias-indicating scores in white people.

As we have seen, the IAT is claimed to predict social behavior, especially in the case of its racial variant and the related measurement of *implicit bias*. Due to the large amount of variants of the IAT, I will make use of this race variant as a case study here. Before we start, a short reminder concerning this variant; a quicker overall reaction time in the white-positive, black-negative task indicates that you have a pro-white (or anti-black) implicit bias, which was dubbed as equal to 'prejudice' in the early publications of the IAT[112]. In the previous chapter, however, we saw there is no reliable effect of this supposed 'implicit bias', or any other type of implicit attitude, on behavior.

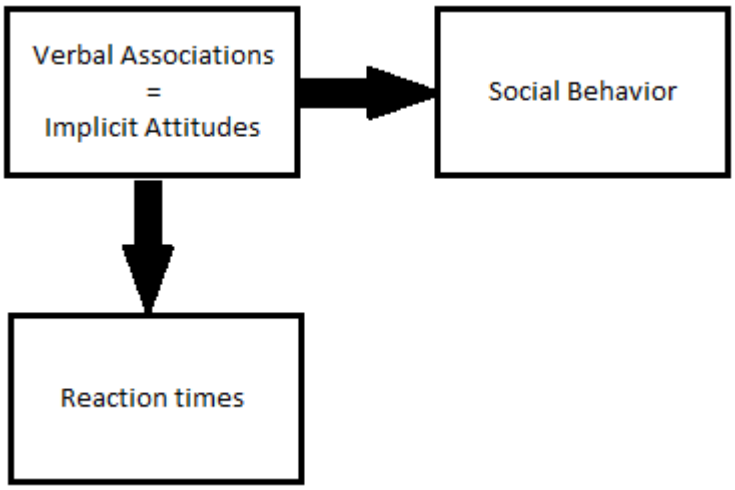

Figure 2: A model of the hypothesized causation structure underlying the IAT

In Figure 2 we see the hypothesized causation structure underlying the IAT[113], which has been described several times before; associations between verbal evaluative categories and verbal

[112] See Schwarz (1998).

[113] This model was based on Greenwald et al. (1998). I have chosen for a 'verbal associations equal implicit attitudes'-variant due to its relative strength over a model in which implicit attitudes influence verbal attitudes, as this is not in line with the claim that implicit attitudes are what is being measured by the IAT.

concept[114] categories cause differences in reaction time, and these same verbal associations are considered to be identical to implicit attitudes[115], or at least, to realize them. These verbal associations/implicit attitudes influence not only our reaction times, but also our social behavior[116]. The direction (i.e., positive or negative) and strength of this influence is dependent on the strength and 'polarity' of the association. The IAT measures the average reaction time difference between two tasks, and thereby picks up on the implicit attitude difference underlying it. The reaction time differential thereby becomes a stand-in for your implicit attitude, just like the speed of the ball became a stand-in for your kicking strength in the earlier example. According to the *implicit social cognition* research program, the reaction time differential can then in turn be used to predict social behavior, such as likelihoods of racial micro-aggressions or other racist behavior.

In order to make this last claim, that we can use IAT results to predict racist behavior, all of the earlier steps need to be true. In the case of a fully verbal IAT, namely one linking verbal evaluative categories with verbal target categories[117], this gives us the following three questions to answer:

1. Can we infer (verbal) associations from reaction time differentials?

2. Can we equate verbal associations to implicit attitudes?

3. Can verbal associations/implicit attitudes predict social behavior?

1. *Can we infer (verbal) associations from reaction time differentials?*

This first question was already mentioned in our discussion of Fiedler's arguments in section 2.8. How logical is the step to verbal associations/implicit attitudes from reaction time differentials? As was discussed there, we cannot be sure that the 'correct' verbal associations are assessed without a model of what these associations are. If *every* association between a concept category and an evaluative category counts as an 'implicit attitude', there might be numerous sub-variants or opposing categories. Only by specifying *which* associations are being measured can we say something about this supposed link, and whether it is possible or not to do so with the IAT. The

[114] A similar structure applies to visual concept - evaluative verbal category IATs and variants like the go/no-go test.
[115] One could argue that by extension these associations become verbal-emotional associations, due to a proposed link between verbal evaluative categories and emotions, or verbal-evaluation associations. These however would introduce several unlikely effects, namely that reading negative/positive words directly affects emotional states or evaluating people worse when a negatively associated word is encountered in conjunction with them, such as vomit, pain or war.
[116] A possible counter-argument would be that implicit attitudes influence both verbal associations and social behavior, but this would leave us with a weaker model, as this would raise the question what 'implicit attitudes' are in this case, as well as how they influence verbal associations and social behavior. Similarly, referring to 'implicit attitudes' as emergent phenomena would leave us with the model proposed here as well.
[117] I.e., without using pictures as a target category or evaluative category. This version requires arguments of itself, due to the difference in hypothesized models - I will not discuss it in the main body of this text due to size constraints and considerations of style.

lack of clarity concerning the specific verbal associations needed however prevents us from doing so.

Inferring verbal associations from reaction time differentials seems to be fine - the thought experiment presented by Greenwald shows this, as does previous research into verbal associations. However, in the case of the IAT, these appear to need to be very *specific* associations. Given Fiedler's arguments, we cannot safely assume that this criterion is met.

2. Can we equate verbal associations to implicit attitudes?

This question poses a problem we have not yet encountered in this thesis. If we assume that the IAT measures the right verbal associations, whatever these may be, we still are left with the question whether verbal associations can be reliably translated into implicit attitudes. In order to answer this question, we need to look at the definition of *implicit attitude* again. Earlier in this thesis, implicit attitude was defined as an *'introspectively unidentified (or inaccurately identified) trace of past experience that mediates favorable or unfavorable feeling, thought, or action toward social objects'*. How does this compare to verbal associations?

While verbal associations can be said to be 'remnants of past experience' like implicit attitudes are hypothesized to be, as you gain associations over time, the main question here is whether verbal associations of any sort can influence our *social behavior, thought or feelings*, as that is the second part of the definition of *implicit attitude*. The possible effects of an *implicit attitude* are stated extremely broadly: as long as it has an *effect* on favorable or unfavorable feeling, favorable or unfavorable thought, or favorable or unfavorable action, and it is an introspectively unidentified remnant of past experience, it is an *implicit attitude*. The notion of implicit attitude therefore *presupposes* an effect on social behavior, of any kind. Therefore, this question can only be answered when we know of effects of verbal associations on behavior.

3. *Can verbal associations/implicit attitudes predict social behavior?*

As we have seen above, we cannot make use of the term 'implicit attitudes' here, as they already *presuppose* an effect on social behavior[118]. Therefore we must limit this question to verbal associations; do verbal associations predict or influence social behavior?

Generally speaking, psychologists know of several associations that have an effect on our

[118] Or social cognition, which in turn would influence social behavior as well.

(automatic) behaviors[119]. Consider Pavlov effects, for example, where a conditioned stimulus starts causing a similar behavioral reaction as an unconditioned, natural stimulus - such as a dog's salivation upon hearing a bell instead of smelling food. An example more closely related to the IAT, due to its use of verbal associations, is the phenomenon known as priming. Priming refers to the activation of a concept or concept category which then influences later behavior, such as word retrieval - when you are primed with 'dog', you are more likely to fill in the blank in c_t with an 'a' than with a 'u', for example. This is an example of *semantic priming;* due to spreading activation, conceptually 'close' words are indirectly activated as well. Some forms of semantic priming can even influence behaviors more conceptually distant from verbal associations than filling in word gaps. A good example is the famous finding that reading words associated with 'old' leads to a lower walking speed[120].

While these examples can be seen as indicating a generalizable link between associations and behavior, there are two large differences between them and the causal model of the IAT, which allow us to make the IAT's problems more explicit. The first difference is technical, while the second is methodological and will be discussed in section 3.5. In all cases of priming or conditioned learning, we see that triggering an association causes a certain behavior; when we ring the bell, the dog salivates, and when we prime someone with word X, we see that they become more likely to use word Y. In the case of the IAT, we assume that something similar happens; the measured association or the implicit attitude will have an effect on behavior[121]. However, this also means that the measured verbal association *must be triggered somehow*. Given the notion that this should happen *implicitly*, without cognitive supervision, we are left with the necessity for humans to *consistently, unknowingly, 'label'*[122] *objects and people in our vicinity*[123] for the IAT's prediction to work, or to at least do so in social situations. Whether this is the case is up for debate. Yet, whatever the conclusion concerning this assumption will be, without confirming it the IAT's proposed model remains unconfirmed and hypothetical.

---

[119] There are of course many types of association-behavior effects. I choose to mention semantic priming and Pavlovian conditioning here as they are the clearest and most established examples. Behavioral priming, for example, is subject to controversy currently, like the IAT, which is why I exclude it here.

[120] See Doyen, Klein, Pichon & Cleeremans (2012), or Bargh, Chen & Burrows (1996). It must be noted that these findings are controversial themselves, however.

[121] A quick counterargument against this would be that the IAT's measured association is not what has an effect, but that this association is instead an indication of a certain pattern of beliefs or more abstract processes that DO influence behavior. This would however change our proposed model, as it adds several new assumptions and steps, complicating it even further and, most importantly, decreasing the chance that the IAT has actual predictive power. This can be inferred from the inclusion of more steps; with every extra necessary step more unrelated factors will influence the process involved, unless it is isolated, which is unlikely given the interconnectedness of brain structures.

[122] I.e., activate verbal associations related with what is perceived.

[123] Versions of the IAT that make use of visual categories are less susceptible to this problem, but still would require spreading, unconscious activation from the visual brain areas to verbal evaluative areas - unless there is a direct link between visual areas and evaluation, in which case the IAT's use of verbal evaluative categories is insufficient and a variant such as the go - no-go test must be used.

Two other assumptions which can be identified in this model are similarly unsupported[124]. First of all, there is the assumption that associations between evaluative words and the activated target category (i.e., the measured verbal associations) should activate *evaluative behaviors and states* about this target category as well. This would necessitate that negatively evaluated words cause negative states, leading to predictions such as that reading negatively evaluated words from the IAT, such as 'war', can cause negative states in isolation of the target category. This claim, too, needs to be proven. Secondly, it must be the case that an activation of the target category also activates related *evaluative* concepts and/or states, without requiring the simultaneous activation of the evaluative category, in order to predict behavior in the real world. In other words, when the target category is activated, it is necessary that either the evaluative verbal category is also activated, and thereby the evaluative state is triggered, or that the target category can directly activate the evaluative category. Within the IAT, both categories are already primed through the first two training sets, which might lead to an exaggeration of the real-world effect. In order to predict social behavior through the IAT, this found association must however also happen outside of this priming.

To conclude our discussion of this third question, we can state that the step from verbal associations/implicit attitudes to social behavior is not proven, as it stands now. The steps from verbal association to influence on an evaluative behavioral reaction could work in a myriad of ways, depending on the hypothesized structure and functionality of associations in the brain. In order to fix this step, several concrete proposals should be made concerning a causal path from verbal associations to social behavior. Next to that, the resulting model should then be grounded in experimental evidence in order to exclude and/or identify possible confounding factors and alternate hypotheses.

A similar answer pertains to the hypothesized IAT model as a whole. In order to take predictions concerning social behavior based on IAT results seriously, we need a justification of these predictions. Without a more detailed model, as we have seen in the three earlier questions, the IAT is reliant on the measured verbal associations *being the 'right associations',* by which I mean the associations which also influence social behavior, and on these verbal associations *having an effect* on social behavior. Neither of these two assumptions seems to be confirmable currently, especially when we consider our conclusion of chapter 2 as well.

This leads us to the conclusion that the *current* model underlying the IAT is insufficient. Questions can be raised concerning every step in the model, except for the claim that verbal associations are measured. This is also, interestingly, the *only* step covered by the thought

[124] The attacks made in this paragraph are partially presented in Greenwald et al. (2002), as the hypothetical model of the IAT.

experiment; we will be faster/slower in the IAT when certain words are more strongly associated. The claim that verbal associations also influence social behavior is completely reliant on – now controversial[125] – empirical observations that those who score as biased against black people also behave as more biased against black people, a claim that is still held by IAT proponents. As we have seen, however, it is unlikely this happens due to *implicit attitudes* which are measured by the IAT. This is the case because the current model rests on several unproven and as of yet unsupported assumptions, such as the resultant need for triggering verbal associations. Next to that, including more ad-hoc assumptions to defend the IAT[126] will only weaken the position that IAT results can predict social behavior, as it theoretically allows for even more confounds and other processes to influence the eventual, resulting behavior.

Due to these problems with the IAT's model, it is in my perspective much more likely that, as Fiedler indirectly argued[127], implicit attitudes measured through the IAT do not really cause biased or racist behavior. Instead, it is *bias and racist behavior which predict IAT scores,* in the sense that actual 'hidden' racists are more likely to gain a bias-indicating IAT score. A racist after all is unlikely to have positive verbal associations with the race he discriminates against, more so than people who do not show racist behavior. This leads to an inflated correlation between IAT results and measurements of biased behavior, without the causation as proposed by the IAT's hypothesized model. While this was the aim of the racial IAT - weeding out 'hidden racists' - it is too sensitive in the way it operates; other 'negative' associations could also lead to a bias-indicating IAT score, without predicting racist behavior.

This alternative explanation explains the relatively low predictive power of the IAT for racist behavior when compared to the grand claims made when it was first published, as well as the large prevalence of racism-indicating results. The low reliability of the IAT, described in section 2.7, however remains unexplained. Yet, this problem can be easily solved by referring to one of the problems with the IAT model mentioned in this section; the results of the IAT are likely dependent on the verbal associations that are triggered by it. These associations, however, might vary. With repeated tests, it is possible that different associations are triggered, dependent on the circumstances, tactics used by the test-taker and other factors. This claim has already been proven in research; when thinking of famous black people known for positive deeds before

[125] See Carlsson & Agerström (2017).
[126] Such as more direct links between associations and behavior, isolated processes or similar explanations.
[127] See Fiedler et al. (2006), pp. 80 - 83. They argue that those with actual biased attitudes are a subset of the bias-indicating IAT scores, and that in this subset the odds of gaining a bias-indicating IAT score is higher than average (due to a higher chance of negative associations with blacks). This would lead to an inflated correlation between bias-indicating IAT results and biased behavior.

taking the IAT, such as Barack Obama or Martin Luther King, we see that IAT scores indicating racial bias can be strongly reduced[128].

3.4 A sociological perspective: II. Neglecting models

We can now state that the IAT was based on several very weak, 'implicit' assumptions. This raises an important question: why was this weakness of assumptions not picked up on in publications for so long[129]?

The literature shows us that, shockingly, Fiedler, Messner and Bluemke's strong arguments have received little to no attention in further literature from IAT proponents, or researchers making use of the IAT[130]. As an example, in section 2.8 we saw that Greenwald and Sriram only reacted to Fiedler's last critique; that a full experiment is necessary in order to prove the IAT. Greenwald and Sriram however neglected to address the critiques of Fiedler which *led up* to this conclusive point, namely that there is no good reason to believe that the IAT measures implicit attitudes at all.

Research that makes use of the IAT after 2006 seems to commit similar mistakes, taking only a small part of Fiedler, Messner and Bluemke's article into account or neglecting its conclusions. For example, consider Falk, Heine, Yuki and Takemura (2009). They cite Fiedler to draw attention to the possible validity problems of the IAT, yet still make use of the IAT paradigm for their research[131]. Similarly, Brand, Heck & Ziegler cited Fiedler in 2013[132], in order to draw attention to the possible 'lack of implicitness' of processes measured by the IAT. Thirdly, in 2016 Van Tuijl et al.[133] cited Fiedler in order to state that the use of cut-off scores indicating bias is perhaps not applicable.

I believe the rhetorical point being made here is clear: Fiedler's arguments have seemingly fallen on deaf ears for the majority of social psychologists. Neither of the last two citations even seems to sufficiently cover the contents of Fiedler's article. Nevertheless, I could extend this list of examples for several more pages, especially if I start including research that does not even cite Fiedler, Messner and Bluemke's article. Usage of the IAT in research has after all continued for 11 years in spite of their criticism.

---

[128] See Blair, Ma & Lenton (2001), and Devine et al. (2012). Given Forscher et al. (2016), these interventions however do not guarantee effectiveness.
[129] This is by far the large majority of social psychology. However, attention must once again be drawn to Fiedler, but also Rothermund and Wentura.
[130] Those critical of the IAT do tend to make use of it; consider Tetlock & Mitchell (2008), for example.
[131] See Falk, Heine, Yuki & Takemura (2009), p. 185.
[132] See Brand, Heck & Ziegler (2013).
[133] See Van Tuijl et al. (2016).

I believe that this lack of interest in, or attention for, a model for the IAT, as was requested by Fiedler, is a symptom of specialization in psychology; the fracturing of psychology into multiple fields with their own relatively narrow specializations. In this section, I will argue that *social psychology* can be seen as a research program with its own assumptions and methods, of which *implicit social cognition* can be seen as a smaller part. In order to do so, I must quickly generalize the field of social psychology. I apologize in advance for possible misgivings or omissions in the description that follows.

Social psychology, as a field of research, specializes in behavior in social contexts, and the various influences on these behaviors, related thoughts and related feelings. It includes, among others the subjects of cognitive biases, stereotyping and emotional reactions. Within the field of social psychology, often a theory is introduced to explain a certain behavioral effect, but the neurophysiological, biological features underlying this effect are considered relatively irrelevant. This is the case because social psychology does not primarily study *how it works*; this is left to other psychologists, such as neuropsychologists, cognitive psychologists and other researchers.

If we identify *social psychology* as a research program, I would point at a *hard core* of belief in unconscious processes, the belief that situational influences have a significant effect on human behavior, and the lack of insight humans exhibit in the causes of their own actions and the fallibility of our own mind[134]. The prevalent methodology in the research program is experimentation, as in almost all of psychology. *Implicit social cognition* can be seen as a smaller research program within the *social psychology research program*, which is even more committed to the existence of implicit processes. Yet, no social psychologist will, as far as I know, argue that implicit processes do not exist, even if they do not subscribe to *implicit social cognition*.

This leads us to the first potential answer to the question: 'why were the internal problems of the IAT's model not picked up on by the entire field of social psychology?' As I have stated above, implicit processes can be argued to be part of the hard core of the *social psychology research program*. This would mean that questioning the existence of these processes would lead to a researcher effectively 'opting out' of social psychology[135]. However, this answer is speculative. We cannot be certain that no social psychologist would 'dare' to argue against implicit processes, nor do I claim to possess a knowledge of the field complete enough to argue that this is the case.

The second possible answer that could be drawn from this description is stronger. This answer points at the apparent lack of research into the physiological mechanisms underlying the

[134] For example, see Kahneman (2011), who summarizes a large amount of social psychological research, and the summary provided in Greenwald & Banaji (1995), pp. 5-6.
[135] Without a belief in implicit processes and the fallibility of human thought, one creates a new position. After all, if no implicit or unconscious processes exist within a theory, all processes must be either conscious, or some third option must be introduced. Also, see Lakatos (1970).

phenomena social psychology studies. As I stated above, social psychology does not consider these phenomena on the level of their implementation (i.e., the physiological structure of the brain) – they leave this to neuropsychologists, cognitive psychologists and other more physically oriented researchers. Instead, they primarily research the effect of inputs (e.g., experimental manipulations, changes in the environment) on outputs (e.g., behavior, choices) of the individual.

In most cases, this lack of interest in physiological mechanisms is not a problem at all; consider priming, for example. As we have seen in the previous section, in the case of priming we observe an effect based on a manipulation; first the priming happens, then effects are observed and compared to those of unprimed participants. In this case, the neurophysiological underpinnings of the effect do not really matter; we manipulate something on the 'outside' of the brain, which leads to a different behavior on the 'outside' of the brain as well. Social psychologists are interested in these *behaviors*, and the *phenomena that influence them*. How these behaviors and influences work exactly in the physiological brain is just not their area; they simply assume that the functions and phenomena they study are somehow realized there.

This type of specialization generally is a good thing. You do not need to simulate brain areas in order to observe behavioral effects following manipulations, nor do you need a model at the neurophysiological level concerning behaviors. In order to observe causations, all you need is to watch inputs and outputs of the individual[136]; what happens when your manipulation is in effect, and what happens if it is not? As an example; if we deprive one of food, water and sleep, this individual eventually dies. If we do allow an individual these things sufficiently, he does not die. To observe this, we do not need to know exactly how the body functions. The same principles apply to behavioral research.

Another common argument in favor of specialization is that, if research into any subject would have to wait until its underlying mechanisms are completely understood, we would likely be forced to stop with all research concerning human behavior for several centuries, or even millennia. We would after all be forced to wait for physicists to draw a definitive conclusion between the string and particle theories of light, wait for cognitive scientists to explain the underlying mechanism of conscious experience, and biologists to understand the understand the apparent intentionality of cellular organelles and RNA.

Yet, in some cases attention must be given to other levels of description, or the 'realizers' of the phenomena one studies. A given theory at a relatively abstract level of explanation, such as cognition, namely tends to indirectly and inadvertently carry assumptions about its underlying levels, which may or may not be true. Take Freud's theories, for example: if we accept

[136] This point will be expanded upon in the next section.

materialism, his theories would require that there are three actively interacting processes in the brain[137], for instance. In the previous section we saw that the IAT also carried such 'hidden' assumptions, in this case about the interaction between verbal associations and behavior: i.e., that such an interaction exists and influences social behavior. Such an interaction however needs to be realized somehow, most likely through the underlying, neurophysiological structure of the brain. This commits those supporting the IAT to the existence of interactions or connections between brain structures realizing verbal associations and those realizing behavior.

However, due to the specialization in and focus on *behavior* social psychologists show, which can be said to be part of the *social psychology research program*, I argue that they have developed a 'blind spot' for such underlying mechanisms of the phenomena they study. As these mechanisms are generally irrelevant for them, they are not used to conducting research on these, nor would they generally pay attention to them. Instead, criticism would be based on a method the researchers *are* comfortable with: experimentation. If you critique or defend a measurement instrument or study, you will after all most likely use techniques you are an expert on, or at least comfortable with, which in the case of social psychologists would be the considerations they put into designing experiments. Examples of these would be identifying possible confounds, testing for replicability, testing various validities and providing alternative explanations, which are necessary tools for experimental psychologists making use of statistical techniques.

With the luxury of hindsight, I can say that this is exactly what happened[138]; the IAT was attacked, but mostly *not* with questions concerning its model or underlying, hidden assumptions. Social psychologists are, after all, *used* to hypothetical models, with unclear physiological substrates, and this is fine in most cases in their field[139] - as long as there is a manipulation involved. The IAT however changed the rules; it hypothesized a plausible cause within the *social psychology* and *implicit social cognition research programs*, namely an implicit factor[140], for observed behavioral effects. However, this cause was also immune to experimentation, as it could not reliably be manipulated, nor could it be directly observed[141]. This immunity was strengthened due to the 'theory-neutral' state of the IAT, which allowed it to remain uncommitted as to its actual functioning[142]. Due to all the above, social psychologists seemingly could only attack the IAT through methodological criticism - which started almost instantaneously, and has not let up over

[137] I.e., id, ego and superego.
[138] See chapter 2 and section 3.1 for examples of these methodological critiques, as well as the beginning of this section.
[139] Consider, once more, priming as an example.
[140] Due to the prevalence of other implicit effects in social psychology, such as priming and heuristics.
[141] See Greenwald & Sriram (2010).
[142] See Greenwald et al. (2005).

the years[143]. From this perspective, we can argue that social psychology *has* been critical towards the IAT - yet seemingly using the wrong angle.

The work of Jan de Houwer of Ghent University[144], a critic of the IAT who has used Fiedler's arguments in his own articles, can serve as a strong support for the claim above. In 2009, De Houwer, Teige-Mocigemba, Spruyt and Moors created a normative analysis of implicit research. Among several other points, they drew attention to the necessity of specifying causal models for implicit measures[145], including the IAT. More precisely, they argued that a 'how' criterion - how attributes causally produce the measurement outcome - must be met in order to support the predictions made by such models. This is very close to the attention to underlying models, which I just accused social psychologists of as missing. Yet, the discussion that followed focused mostly on hypotheses concerning processes causing IAT effects, which resided on a similarly abstract level as the standing hypothetical model of the IAT[146]. Next to the 'how' criterion, they argued in favor of a 'what' criterion - which attributes causally produce the measurement outcome. In their discussion of this 'what' criterion in relation to the IAT, they mostly focused on the possible confounds of the IAT within the current model, and the IAT's statistical properties, such as its predictive validity[147]. This illustrates the 'social psychological blind spot' I have introduced in this section; it could equally well be argued to be *too much attention* to the experimental and statistical properties of theories. However, this attention for one aspect of the phenomena they study also seems to blind them to others, in this case the hidden, unproven assumptions of the IAT[148].

In conclusion, we can state that the IAT *was* scrutinized and criticized - extensively, even, given what we have seen in chapter 2 and this section. However, due to their specialization, social psychologists were very unlikely to pick up on the logical and methodological problems of the IAT. Then again, we must add that these problems *were* picked up on, at the very least by Klaus Fiedler and his co-authors. Yet, apparently these people were not with enough to force a stop on using the IAT until its assumptions could be proven.

---

143 See chapter 2. The covered timeline of critiques and refutations of critiques on the IAT ranges from articles from 2000 (Dasgupta et al.) to 2016 (Carlsson & Agerström).

144 Critical, because he published an alternative explanation in 2005, and effectively is granted his own paragraph in Teige-Mocigemba et al. (2010).

145 See De Houwer, Teige-Mocigemba, Spruyt & Moors (2009).

146 See De Houwer et al. (2009), pp. 354 - 356.

147 See De Houwer et al. (2009), pp. 351 - 354.

148 As a contrast to the 'social psychological blind spot', I wish to draw attention to the work of David Amodio, a 'social' neuroscientist who introduced a neurology-based framework for implicit social cognition, the Multiple Systems Model (MSM). This model introduces three different 'memory systems' instead of one 'association system' as an explanation of implicit social cognition, all three of which are directly linked to neurophysiological areas. See Amodio & Ratner (2011).

3.5 A logical perspective: II. Black-box thinking and concluding causes from effects

At this point in this thesis, we can readily state that the IAT was build on several problematic assumptions which are not strong enough to uphold the claims made in its name, and of which the researchers seem to have been unaware[149]. This last fact can be gauged from the lack of research into the underlying claims of the IAT, the motivation of which can be seen in the arguments Greenwald and Sriram[150] gave to defuse the attacks on the IAT; maybe we cannot test for causation, but correlation is enough. Yet, this claim would only be warranted if the underlying model was strong enough to 'carry' their claims, however, as we saw in section 3.2. Next to that, due to the weakness of its underlying assumptions and the lack of a model, the IAT cannot be said to predict anything at all. In this section, I will discuss the logic underlying the interpretation of the IAT when the assumptions discussed in section 3.3 are eliminated, expanding on the discussion of abduction given in section 3.2.

In section 3.4, I stated that social psychologists primarily watch behavioral inputs and behavioral outputs. This method allows for a technique which is known from experimental behaviorism; black-box thinking. Black-box thinking refers to the treatment of a certain object as a black box, which means that the internal functioning of this object is unknown, yet we can still observe what enters and exits the box. A computer can be treated as a black box for example; we can see what we type and what happens on the screen, but we don't need to know how the computer works to deduce a causation between our typing and the words appearing on the screen. In behaviorism, the object treated as a black box is the brain of the individual, while one could, for example, also treat a car engine as a black box for driving purposes.

In any case of black-box thinking, if pressed, a hypothetical explanation can be offered as to what the black box 'does', without needing to discuss the internal structure of the black box. In behaviorism, this happens through the use of dispositions combined with motivational states, which basically restate the input and output in terms of basic behavior: 'If one is thirsty, he has the disposition to go to the tap and drink water'. I call this *'labelling the black box'*, as no real explanation is given for what happens inside the black box[151]. Instead, only a *hypothesis* of the internal functioning of the black box is given, which is almost completely reducible to the cause

[149] As they have - as far as I could find - not mentioned them in articles, nor have they published research about them.

[150] See Greenwald & Sriram (2010)

[151] This is similar to Dennett's virtus dormitiva, but more specifically linked to a mechanism. Instead of simply citing him and linking his work to my point here, I have chosen to stay with my own terminology in order to make it more accessible for those unaccustomed with his work, whilst also avoiding possible mistakes in interpretation or 'strong-arming' his philosophy into my mold. Nevertheless, I am indebted to him for this insight. See Dennett (1978), 'Skinner Skinned' for this argument.

and effect themselves[152]. In effect, what happens is after all equal to putting a label on the black box which says *'Contains the X-er'*, in which X stands for the action performed by the black box, essentially explaining a function with a hypothetical something that performs that function[153]. While this could be seen as a form of abduction - as a cause is inferred to account for an observed phenomenon - it does not provide an actual explanation at all, unlike the abduction we have seen so far. Instead, we explain an observed effect by a hypothetical 'causer', a mechanism which is unobservable.

Black-box thinking, like abduction, is not necessarily wrong, however. The value of black-box thinking rests in its affordance of making claims based on observed inputs and effects *without* needing an underlying physical explanation or an implementation-level model. You see what goes in, and then you see what goes out. If you press the button, the corresponding symbol appears on the screen. This allows us to create laws and make predictions *without* needing to understand the entire mechanism underlying them.[154]

In the case of the IAT, black-box thinking seems to have been applied. Greenwald has, after all, stated that the IAT is 'theory-neutral'[155], meaning that the exact process underlying the IAT is to be determined, yet is not relevant for the overall effects observed in the paradigm. Next to that, the definition for *implicit attitude* is quite vague; *'an introspectively unidentified (or inaccurately identified) trace of past experience that mediates favorable or unfavorable feeling, thought, or action toward social objects'*. This allows for any neural connection or neuron of some sort[156], which is connected to an area that is involved in social behavior, feelings or thoughts, to be 'an implicit attitude', making them impossible to track down or pinpoint, and allowing for any number of ad hoc hypotheses concerning their functioning or implementation. The actual physical substrate is irrelevant for the definition; only *the effect it has* is relevant, besides the implicitness. An '*implicit attitude'* can therefore be equated to an *'implicit-learned-social-behavior-influencer',* leading to the conclusion that it is nothing more *than a label on the black box*. This lack of theory surrounding the IAT effectively makes it immune to criticism, allowing most attacks to be parried with 'but implicit attitudes might work differently' through pointing at the 'uncommitted' nature of *implicit social cognition*.

In general, black-box thinking is not too much of a problem, as we have seen in our

---

[152] It must be added that social psychologists *do* often add explanations further than simply labelling the black-box, but rarely on an implementational level such as brain functioning, which would be the internal structure of the brain. Instead they rely on abstract descriptions of what happens in the black box, which makes it a more sophisticated form of labelling.

[153] Another example would be inferring the existence of a thunder-god from thunder: thunder happens due to the 'thing that causes thunder'.

[154] Again, this can be linked to Dennett, this time to his three stances. My reasoning behind not citing him is idem to footnote 150. See Dennett (1987).

[155] See Greenwald et al. (2005).

[156] As every 'trace of past experience' in the brain can only be implemented as a change in neural connections or the activation threshold of a (set of) neuron(s).

example with the symbols appearing on the computer screen. Another unproblematic example would be the knowledge you need to *drive* a car. You will need to know that pressing the gas pedal will move the car forward, and that braking will make it slow down. Similarly, you will need to know that the car must be in gear and with the engine turned on to function. How the engine works exactly does not really matter; these basic principles are correct for all cars, no matter whether they have a standard four-stroke engine, a two-stroke engine or an electric engine. However, in the case that your engine breaks down, or when you claim something about the engine itself, you *do* need to have knowledge about car engines to be able to say more than 'the engine broke down' and to be able to repair it. You can not only rely on the *external* causes and effects surrounding the black box and your observed 'laws', you also need to know the *internal* causes and effects, namely how the engine functions and what parts and processes are responsible for which output effects[157].

In the case of the IAT, making use of black-box thinking however was a large mistake. Where normally black-box thinking suffices for social psychology, such as in the case of priming, here it does not. Why? Simply put, because claims were made about completely *internal* causes and effects, things that happen *inside the black box* of the brain, without caring about the *internal structure* of the brain[158] nor *observing an input*. In the case of priming, we could instead use the observable manipulation to determine a causation.

The IAT relies on the existence of *implicit attitudes*, unobservables from a first person perspective which nevertheless exist somewhere, someway in the brain, which influence both reaction times and social behavior. These implicit attitudes were inferred from the possibility of unconscious influences on social behavior, and their link with verbal associations seems to have been a product of the creation of the IAT. This leaves us with *implicit attitudes* as a *hypothesized* cause, which is supported by looking at two forms of *output* of the 'black box', the brain in this case; reaction time differentials and social behavior. No real input can be observed[159], nor can we look 'inside the black box'; we do not know what an *implicit attitude* actually is, nor what it consists of or how it works. This allows us to question a very basic assumption of the *implicit social cognition* research program; *do implicit attitudes actually exist?* Is there even something that has an effect on

[157] In literature this has been described as the difference between etiological and mechanistic explanations, where etiological explanations only describe causes and effects through a law, while mechanistic explanations also describe the process underlying this law. See Craver & Tabery (2015) for more information on this distinction.
[158] I'm assuming a materialist position in the mind-body debate here, which might be attacked. Yet, when exchanging 'brain' with 'mind', the problem described still holds.
[159] Of course, Greenwald & Banaji (1998) have proposed a cause for the creation of implicit cognitive effects; prior experience. However, this is not exactly measurable in an experiment.

both verbal associations and social behaviors, and is *one singular, measurable cause*?[160]

There is only one way to answer these questions. In the case of the IAT, an internal cause, the implicit attitude, is inferred from two different outputs. This internal cause explains both of the outputs, but, as we have seen, in effect is nothing more than a label for *'something that causes both outputs'* without an underlying explanation. Translated to our example surrounding the car, this is similar to predicting that a car moves forward and that its brake lights are off by using the explanans *'an unobservable something in the car that causes brake lights to be off AND the car to move forward'*, which can be mirrored by a definition of implicit attitudes as *'an implicit something*[161] *that has an effect on social behavior, thoughts or feelings AND reaction times in the IAT*'. However, in order to make the claim that two effects have a shared cause, you will need to know *how* these effects come to be, and whether they are actually interrelated in the internal structure, as we have seen in the previous section. Next to that, you also will have to pinpoint what the proposed shared cause *is*, within this internal structure.

As we have seen in section 2.7 and 2.8, proponents of the IAT have however avoided doing *exactly these things;* they have not created a model, nor have they done research into the 'how' of implicit attitudes' effect on social behavior[162]. Indirectly, we can gather that the *implicit attitudes* measured by the IAT are realized through verbal associations. But which verbal associations exactly? And how does their proposed vision of 'verbal associations' work in the brain, influencing both social behavior and reaction times without being consciously accessible? Is there a direct link between these things, or does it work more obliquely, through the verbal associations affecting the holistic, overall brain state indirectly and thereby impacting social behavior? These questions have remained unanswered, yet they are *necessary* for the claim made by IAT proponents; that implicit attitudes cause both reaction time differentials and social behavior.

A counter-argument to the argument which I propose in this section could be that reaction time differentials measure implicit attitudes/verbal associations extremely reliably. This simplifies the model, as we can now say that we are reliably measuring a phenomenon in the black box and do not have to worry about the step from the reaction time differentials to implicit attitudes anymore. However, this would still require an *internal* explanation for the proposed link

---

[160] Machery (2016) makes an interesting point that an 'implicit attitude', in the sense of prejudice, is actually only possible as a *trait*, which emerges from multiple different brain areas. According to him, the IAT measures just one of these, and therefore cannot be said to measure 'prejudice' at all.

[161] A 'trace of past experience' after all is non-informative when talking about the brain, since all neural connections can be said to be either 'a trace of past experience' (i.e., learned) or part of a necessarily rigid network structure (such as V1).

[162] I am not counting the model of Greenwald et al. (2002), since it can be attacked by the same arguments as those presented here - and several more that have not been discussed yet, such as their simplified vision of negative and positive as singular conceptual categories and the idea that single concepts can be treated as nodes in a connectionist network.

between verbal associations and social behavior on an implementation level, and would necessitate the measurement of 'the correct verbal associations' as was mentioned in the previous section. Similarly, this counter-argument would allow for direct experimental tests of the IAT model, by measuring changes in social behavior after retraining verbal associations in both racism-inducing and racism-reducing directions. However, this possibility has been denied by Greenwald and Sriram, when they stated that one cannot measure whether the implicit attitudes have actually changed[163]. This indicates that the measurement of the reaction time differentials is not seen as a completely reliable measurement of verbal associations/implicit attitudes by the proponents themselves.

To conclude, we can state that the IAT is commonly interpreted by making use of severely flawed logic; inferring an *effect/output*, social behavior, from a hypothetical *cause*, implicit attitudes, that was inferred from another output, namely reaction time differentials, without sufficient reasons[164] to tie these together. As simple as it seems, this mistake is quite grave, and seemingly has not been picked up on in almost 20 years of research on the IAT. In my perspective, this has happened due to an interpretative misstep, namely the intuition that someone who has more negative verbal associations with one race over the other is racist – and therefore will act like a racist as well. While this sounds *intuitively* logical, it is dependent on the nature of negative verbal associations and their influence on behavior[165]. Explanations on a neurophysiological level are required to solidify this intuition. However, instead of researching such an explanation, the notion of *implicit attitude* was used a bridge between verbal associations and behavior, but in fact this remained *'labelling the black box'* as no mechanism was introduced to explain the proposed causation. Meanwhile, all proof that pointed into the direction predicted by *implicit attitudes* was interpreted as support of their existence.

3.6 Three perspectives: A summary

In this chapter, I discussed the IAT from sociological, methodological and logical perspectives, and have shown that several mistakes were made in the scientific process underlying it. Firstly,

---

[163] See Greenwald & Sriram (2010), p. 238. In comparison to the first introduction of implicit attitudes, in Greenwald & Banaji (1998), the process responsible for implicit attitudes is described vaguely at best as well; "*The implicit processes conceived in the present analysis are, in part, subsumed by the notions of peripheral or heuristic processing, but also involve processes operating even further from the range of conscious thought than conceived in these analyses.*" (p. 5).

[164] What reasons would count as sufficient is a large debate in philosophy of science, which I will not discuss here due to concerns with textual focus. For the present discussion, I consider a confirmed implementation level model (i.e., describing the mechanism underlying the effect) or a successful experimental manipulation (such as in the case of priming) as sufficient for drawing the intended conclusions, as I have argued earlier in this thesis. Also see Chapter 4 for a continuation of this discussion.

[165] See section 3.3 for a more detailed discussion of this point.

from the methodological perspective, we have seen that the hypothesized causal model of the IAT is weak, and that it relies on several unconfirmed and unlikely assumptions concerning the nature of verbal associations and the brain. Secondly, from the logical perspective, we have seen that the IAT's interpretation relies on weak abduction, and a hypothetical shared cause for two output effects, implicit attitudes, which remains unconfirmed to date. Thirdly, from the sociological perspective, we have seen that the overzealous interpretation of the IAT by its creators can be seen as influenced by the *implicit social cognition* research program, while the lack of attention directed to the underlying model of the IAT can be seen as a symptom of specialization in psychology, which has in turn lead to a lack of attention to underlying mechanisms in the field of social psychology.

4. Diagnosing the IAT Controversy: Conclusion

In this thesis I set out to perform a case study on the IAT controversy aimed at identifying its possible causes, besides those known from the ongoing discussion of the replication crisis in social psychology[166]. In this chapter, I give a final 'diagnosis'[167] of the IAT controversy, ending this case study. I do so by combining the problems identified in chapter 2 and 3, into two lines of argumentation: one concerning the scientific process, focusing on the methodological and interpretation problems of the IAT, and one concerning the possible sociological causes of the mistakes in the scientific process, focusing on the circumstances in which these occurred. A brief discussion of the recommendations we can draw from this case study is included at the end of this chapter.

4.1 Diagnosing the IAT controversy: Refreshing our memory

Before we can start our diagnosis, it is helpful to refresh our knowledge of chapter 2's conclusion. Firstly, we saw that the IAT was created from the perspective of a research program, *the implicit social cognition research program*, with the aim of measuring implicit attitudes. We then saw that, from within the field of psychology, severe criticism has been launched against the IAT, primarily focusing on its test-retest validity and predictive validity, both of which can be considered low or even non-existent. Following this was a summary of the methodological critique of the IAT by Fiedler, Messner and Bluemke[168], which pointed out several failings in the underlying model of the IAT. In conclusion, we were able to state that the current interpretation of the IAT, namely that it measures implicit attitudes which can predict social behavior, is unsupported by the empirical evidence.

4.2 Diagnosing the IAT controversy: The scientific process

We can now turn to our diagnosis of the scientific process underlying the IAT controversy. Yet, before we start a quick definition of the term 'scientific process' might be helpful. By 'scientific process underlying the IAT controversy', I refer not only to the creation process behind the IAT,

[166] For instance, questionable research practices and publication bias. See
[167] In this chapter I will use the words 'diagnosing' and 'diagnosis' as a descriptive metaphor, in the sense that I identify various 'symptoms' and try to link these to possible causes. This metaphor continues throughout this chapter.
[168] See Fiedler et al. (2006).

but also to the treatment of the IAT in scientific literature, the critiques that were aimed at it and the defenses raised against those critiques, as well as the contents of these critiques. Most of these points have already been described in chapter 2 and 3, where we concluded that the IAT has severe problems with its underlying model and empirical support.

The core problem with the scientific process which we can distill from these earlier discussions, is that the claims made at the IAT's conception, namely that it can measure implicit attitudes and predict social behavior, have never been justified. There is no strong, stable or proven effect of the hypothesized *'implicit attitudes'* on racist behavior, nor is there sufficient support to believe that *implicit attitudes* can be reduced to simple association strengths, or that these association strengths are influenced by *implicit attitudes*, or realized through them. *Implicit attitudes* themselves seem to only be 'X'ers' in the black box brain, a point that is strengthened by their 'theory-neutral' background. Next to that, they are a hypothesized cause for two distinct types of effects, without a model to bind them all together. This makes the current 'orthodox' interpretation of the IAT, at the very least, no better than some of its alternatives, and raises the question why this interpretation was chosen to begin with: why did the creators of the IAT abduce towards this single interpretation? Why was it 'the best'?

A justification is necessary in order to make the current interpretation (i.e., the IAT measures implicit attitudes, which can predict social behavior) more than 'an inference to the *first* explanation', to change it from an unconfirmed hypothesis into a theory with empirical foundations. In any scientific environment, *proof* is required for a claim to causation, such as Greenwald and Banaji have made in their 1998 press release[169]. Without it, the unsupported interpretation is just a wild hypothesis, not a basis for further research, and especially not a candidate for real-life implementation, as has already happened with the IAT. Yet, the main proof presented for the current interpretation of the IAT (i.e. that it reliably predicts racist behavior) is voided in the light of Carlsson and Agerström's meta-analysis[170] and the meta-analysis by Forscher et al.[171].

One can easily point out that this rebuttal has been published many years after the IAT's publication. Yet, the current interpretation of the IAT was published without the empirical evidence that would eventually come to support the IAT in the 00's. Even in the actual article introducing the IAT, Greenwald, McGhee and Schwartz were hesitant to make a large claim to causation.

The current interpretation of the IAT is, and was, therefore 'running on fumes'. The only

[169] See Schwarz (1998).
[170] See Carlsson & Agerström (2016).
[171] See Forscher et al. (2016).

thing differentiating it from alternative explanations are differing hidden, or, '*implicit*' assumptions about the inner workings of the brain and mind, which are untestable as of yet; a supposed link or mechanism acting between verbal association, evaluation and action, as was discussed in section 3.3. Without an explanation why we should believe the IAT's proponents and their interpretation over the countless other possible interpretations of the current body of IAT results, there is however no reason to subscribe to their viewpoint, except for personal or environmental ones. This marks our move away from purely scientific discussion and a more sociological approach. In the next section, I will return to this in more detail.

To conclude, we can state that the IAT controversy was in part caused by a lack of methodological and logical rigor. This problem is twofold. First of all, the causation predicted by the IAT's proponents, namely that implicit attitudes influence social behavior and affect IAT results, remains unsupported to date. The cause of this lies mainly with a lack of attention to the underlying, implicit assumptions of the IAT, which were discussed in section 3.3. Secondly, there was little to no reason to abduce towards the current interpretation instead of the other possibilities to begin with. Yet, the claim that the IAT could be used to predict racism was still made.

4.3 Diagnosing the IAT controversy: The sociological causes

Besides problems located in logical and methodological areas, possible sociological causes for the IAT controversy have also been discussed. Most notably, I have argued that *implicit social cognition* and *social psychology* can be seen as research programs, and that they both have influenced the IAT controversy. Firstly, I argued that the *implicit social cognition* research program has defended itself following Lakatos' ideas; by making use of the positive heuristic and by defending the hard core. Secondly, I have argued that *social psychology* as a specialized research program seems to have created a 'blind spot' for the underlying (neurophysiological) assumptions of theories in its proponents. In this section, I propose a possible answer to the question why the problems with the IAT do not seem to have been picked up on.

The problems of social psychology can be argued to be applicable to those within the implicit social cognition research program as well; the creators of the IAT seem to have not paid too much attention to the underlying (implicit) assumptions of their own theory, just like most of their critics and supporters. Given the 1995 article introducing the possibility of *implicit social cognition*, it is also likely that Banaji and Greenwald were highly excited by the results they found when they first tested the IAT, namely results that supported their earlier predictions. Building

from this, we can argue they applied the 'positive heuristic' to the IAT. They did search for alternative explanations, even mentioning them in their introductory article, but nevertheless would focus mostly on the results reinforcing their theory. This lowers the chance that 'sufficient' scrutiny (i.e., scrutiny that would lead to the discovery of the issues discussed in the previous section) would come from the creators of the IAT. A last argument can be drawn from the *theory-neutral* state of the IAT, which was used by Greenwald in his defense against Rothermund & Wentura, and Friese & Fiedler[172]. This defense, namely that the IAT is not specifically committed to a certain type of underlying mechanism, also defuses the possibility of attacking the IAT through any form of mechanistic criticism.

Similarly, we can argue that enthusiasm and the positive heuristic has led to the unjustified abduction Greenwald and Banaji made: they inferred to this explanation because their first found results were in line with their predictions. According to their research program, their claims were soon to be completely verified, while this research program itself was based on several decades of previous research. They did not abduce into the causal relationship between implicit attitudes and social behavior out of nowhere: they believed they had sufficient proof, and that their interpretation was the most likely. One could also invoke the social psychological notion of *confirmation bias* here; the bias to interpret information in such a way that it conforms to what you already believe.

Next to the sociological factors surrounding the IAT's creators, more can be said about the missing scrutiny from inside the field of social psychology. As I argued in section 3.5, researchers in the *social psychology research program* are not used to discussing mechanisms underlying the phenomena they research , as these can generally be considered part of the neuropsychological domain. Instead, they focus on behaviors, effects and manipulations, and use experiments to research these – somewhat mirroring the 1950's behaviorism. This can be added to the fact that the IAT was most likely to be attacked from within the field of social psychology, as it is based on findings from the field of social psychology as well as being a prominent measurement instrument within this field. This combination of facts decreases the likelihood that a lot of attention would be directed towards the IAT's assumptions from outside the field of social psychology, whilst the field of social psychology itself has already been shown to be relatively 'oblivious' to these.

Another possible cause can be found in the introduction; *the pressure to publish*. If researchers have to fight for publication in order to keep their jobs, it is likely that fundamental questions which are not in the area of expertise of individual researchers will remain

[172] See Greenwald et al. (2005).

unquestioned, such as in this case the question concerning the physical mechanism of *implicit attitudes*. Instead, they will focus on research questions that provide 'an easy way to score', choosing questions that will quickly provide a publication. Usage of the IAT, especially in its racial bias variant, would nearly guarantee this due to its large possibility of significant effects, because studies with significant results have a higher chance of publication. Similarly, the question into the physical mechanism underlying the IAT is not even near the area of expertise of a social psychologist, and most likely out of reach for the current state of neuropsychology as well. Due to the vagueness of *implicit attitudes*, it is also nearly impossible to deny their existence through experimentation; they can be realized through anything that has an implicit effect on social behavior. This further decreases the odds that critiques would concern the IAT's model instead of its statistical properties.

In conclusion, I argue that sociological factors influenced both the adoption of the current interpretation of the IAT and the (lack of) criticism that followed it. Nevertheless, attention must be drawn to the fact that several researchers, including Fiedler, Tetlock and Mitchell[173], have been critical from the beginning of the IAT's 'cult status'. Similarly, it must also be noted that the proponents of the IAT have not made condemning mistakes either – at least, from a Lakatosian perspective. After all, so far they have applied the positive heuristic, and no completely damning evidence has surfaced that shows that the current interpretation of the IAT is completely and utterly wrong. The meta-analyses by Oswald[174] and Carlsson[175] show mainly that the predicted effects are, respectively, not as strong as thought before, or not entirely proven due to methodological mistakes. The meta-analysis by Forscher[176] might show that changes in IAT results do not lead to significant changes in behavior, yet doesn't

4.4 Diagnosing the IAT: Final conclusion and recommendations

With these two discussions, I have fulfilled the aim of this thesis: diagnosing the scientific process underlying the IAT controversy. As a final conclusion, I argue that problems on a methodological and logical level were ignored, such as the lack of an underlying model supporting the IAT's predictions, and a (too) quick abduction towards the current interpretation of the IAT. A likely cause for these problems lies with several sociological factors surrounding the IAT, such as the earlier *implicit social cognition* research program and a possible blind spot of social psychologists for

[173] Tetlock and Mitchell have contributed in Oswald's meta-analysis for example, but also co-authored the critical Tetlock & Mitchell (2008).
[174] See Oswald et al. (2013).
[175] See Carlsson & Agerström (2016).
[176] See Forscher et al. (2016).

underlying methods.

The primary focus of this thesis has been the diagnosis of the IAT controversy. Yet, from our diagnosis we can also draw several recommendations for the future. Some possible recommendations are fairly obvious; that new claims about causation should be made with caution until the underlying model is verified, or that it is necessary to prove your theory before building interventions on it outside of academia. These recommendations are straight-forward, and already quite well-known maxims within the scientific community. The following two recommendations are less obvious.

Firstly, we can conclude that attention and research should be directed to the underlying model of the IAT if progress is to be made regarding the claims of both IAT proponents and opponents. How would implicit attitudes exert influence in the world? And how are they realized? More concrete proposals towards the physical or neurophysiological properties of the IAT's conceptual background, especially *implicit attitudes*, must be made. As we have seen, the current interpretaton of the IAT's properties is after all not falsified. Instead, we can say that its foundations have been severely shaken, and that now, it is time to either attempt to rebuild[177] or to abandon them. In order to do the former, I believe it is necessary to justify the abduction towards *implicit attitudes*, and that the aforementioned focus on underlying models and assumptions will assist in this.

Secondly, if social psychologists do not already do so[178], I also wish to advise them to pay (more) attention to the assumptions about the real-world their theories and testing paradigms carry: their 'implicit assumptions'. While a hypothesized cause-effect relationship can be proven through good experimentation only, a hypothesized model also requires proof for its internal mechanisms. In this thesis I hope to have shown that it is not enough to give a plausible account of what is happening if large claims are to be made, such as in the current interpretation of the IAT. Similarly, I hope this thesis shows that fundamental research into the underlying assumptions of the model is, at the very least, useful for disproving alternative explanations, and, at most, necessary for dispelling doubt about a theory.

[177] Attempt, because it is possible that in the process of rebuilding, the original interpretation must be changed.
[178] Given the large amount of criticism the IAT received, one could argue that they already do.

5. Afterword

In the introduction, the police officer seemed to be a 'victim' of several researchers making overenthusiastic claims, together with the government that sponsored his training. However, I hope that readers of this thesis will not walk away with the belief that there were *perpetrators* behind the IAT controversy*,* in the sense that the IAT could be seen as a conscious scam. While blame could be directed towards the creators of the IAT for the controversy – as they actively argued that the IAT was theory-neutral, didn't see the necessity of theory to support their abduction, and made the grand claim that the IAT could be used to predict racism – I do not believe they were aware of the problems pointed out in this thesis, nor that they consciously hid these.

I believe that the IAT controversy should in the future be seen as a methodological scarecrow. It is a good example of what can go wrong when underlying models are ignored, and when discussions focus too little on the reality at stake. The lack of theory and modeling surrounding the IAT, in combination with the proclaimed 'theory-neutral' background effectively made the IAT immune to criticism. At the same time, social psychology's blind spot for these properties of the IAT allowed for the discussion to continue for years.

While the causes identified in this thesis can be seen as an addition to the known reasons for the replication crisis, caution must be advised. Given that this was a case study, it is possible that identified issues with the scientific process, such as the lack of consideration of alternative hypotheses, do not extend to other theories, claims or methods under scrutiny. Nevertheless, this problem may have occurred elsewhere as well. Therefore I do wish to advise other researchers of the replication crisis to take the causes identified in this case study into account.

As a final word; I hope that this thesis provides a good start for learning from the mistakes made in the IAT controversy, so that in the future attention will be given to the possible 'implicit' assumptions underlying models.

6. References

Abdollah, T. (2015, March 9). Police agencies line up to learn about unconscious bias. *Police One,* retrieved from https://www.policeone.com/patrol-issues/articles/8415353-Police-agencies-line-up-to-learn-about-unconscious-bias/

Amodio, D.M. & Ratner, K.G. (2011). A memory systems model of implicit social cognition. *Current Directions in Social Psychological Science,* 20-3, pp. 143 - 148. doi: 10.1177/0963721411408562

Andreychik, M. & Gill, M.J. (2012). Do negative implicit associations indicate negative attitudes? Social explanations moderate whether ostensible "negative" associations are prejudice-based or empathy-based. *Journal of Experimental Social Psychology,* 48, pp. 1082 –1093. doi: 10.1016/j.jesp.2012.05.006.

"Are you prejudiced? Take the Implicit Association Test" (2016, March 7). *The Guardian*, retrieved from https://www.theguardian.com/lifeandstyle/2009/mar/07/implicit-association-test

Bakker, M., van Dijk, A. & Wicherts, J.M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science,* 7-6, pp. 543 - 554. doi: 10.1177/1745691612459060

Bargh, J.A., Chen, M. & Burrows, L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype-activation on action. *Journal of Personal and Social Psychology*, 71-2, pp. 230 - 244.

Bassili, J.N. (1996). Meta-judgmental versus operative indexes of psychological attributes: the case of measures of attitude strength. *Journal of Personality and Social Psychology*, 71-4, pp. 637 - 653. http://dx.doi.org/10.1037/0022-3514.71.4.637

Baumeister, R.F. & Tierney, J. (2012). *Willpower, Rediscovering the Greatest Human Strength*. London, UK: Penguin Press

Beres, D. (2016, October 12). Yes, Implicit Bias Exists - No, That Doesn't Make You Racist. *BigThink*, retrieved from http://bigthink.com/21st-century-spirituality/implicit-bias-is-not-racism

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

Blanton, H., Jaccard, J., Klick, J., Mellers, B., Mitchell, M. & Tetlock, P.E. (2009). Strong Claims and Weak Evidence: Reassessing the Predictive Validity of the IAT. *Journal of Applied Psychology*, 94-3, pp. 567 - 582. doi: 10.1037/a0014665

Brand, R., Heck, P. & Ziegler, M. (2014). Illegal performance enhancing drugs and doping in sport: picture-based brief implicit association test for measuring athlete's attitudes. *Substance Abuse Treatment, Prevention, and Policy*, 9-7. https://doi.org/10.1186/1747-597X-9-7

Brownstein, M.S. (2015). Implicit Bias. *Stanford Encyclopedia of Psychology*, retrieved from https://plato.stanford.edu/entries/implicit-bias

Carlsson, R. & Agerström, J. (2016). A closer look at the discrimination outcomes in the IAT literature. *Scandinavian Journal of Psychology*, 57, pp. 278 - 287. doi: 10.1111/sjop.12288

Cesario, J. (2014). Priming, Replication, and the Hardest Science. *Perspectives on Psychological Science,* 9-1, pp. 40 - 48. doi: 10.1177/1745691613513470

Craver, C. & Tabery, J. (2015). Mechanisms in Science. *Stanford Encyclopedia of Philosophy*, retrieved from https://plato.stanford.edu/entries/science-mechanisms/

Curate Science (n.d.). Retrieved March 2nd, 2017 from http://www.curatescience.org/#ego-depletion

Cvencek, D., Greenwald, A.G., Brown, A.S., Gray, N.S. & Snowden, J.S. (2010). Faking of the Implicit Association Test is Statistically Detectable and Partly Correctable. *Basic and Applied Social Psychology,* 32, pp. 302 - 314. doi: 10.1080/01973533.2010.519236

Dasgupta, N., McGhee, D.E., Greenwald, A.G. & Banaji, M.R. (2000). Automatic Preference for White Americans: Eliminating the Familiarity Explanation. *Journal of Experimental Social Psychology*, 36, pp. 316 - 328. doi: 10.1006/jesp.1999.1418

De Houwer, J., Teice-Mocigemba, S., Spruyt, A. & Moors, A. (2009). Implicit Measures: A Normative Analysis and Review. *Psychological Bulletin*, 135-3, pp. 347 - 368. doi: 10.1037/a0014211

Dennett, D.C. (1978). Skinner Skinned. In Dennett, D.C. (Ed.), *Brainstorms: Philosophical Essays on Mind and Psychology* (pp. 53 - 70). Cambridge, MA: MIT Press

Dennett, D.C. (1987). Three Kinds of Intentional Psychology. In Dennett, D.C., *The Intentional Stance* (pp. 43 - 68). Cambridge, MA: MIT Press

Devine, P.G. (1989). Stereotypes and Prejudice: Their Automatic and Controlled Components. *Journal of Personality and Social Psychology,* 56-1, pp. 5 - 18.

Devine, P.G., Forscher, P.S., Austin, A.J. & Cox, W.T.L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, 48-6, pp. 1267 -1278. doi: 10.1016/j.jesp.2012.06.003

Douven, I. (2017). Abduction. *Stanford Encyclopedia of Philosophy,* retrieved from https://plato.stanford.edu/entries/abduction/

Dovidio, J.F. & Gaertner, S.L. (1986). Aversive Racism. *Advances in Experimental Social Psychology*, 36, pp. 1 - 52.

Dovidio, J.F., Kawakami, K., Johnson, C., Johnson, B. & Howard, A. (1997). On the Nature of Prejudice: Automatic and Controlled Processes. *Journal of Experimental Social Psychology*, 33-5, pp. 510 - 540.

Doyen, S., Klein, O., Pichon, C.L. & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, But Whose Mind? *PLOS One*, 7-1, e29081. https://doi.org/10.1371/journal.pone.0029081

Earp, B.D. & Trafimow, E. (2015). Replication, falsification and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6-621, retrieved from http://journal.frontiersin.org/article/10.3389/fpsyg.2015.00621/full#h5. doi: 10.3389/fpsyg.2015.00621

Falk, C.F., Heine, S.J., Masaki, Y. & Takemura, K. (2009). Why Do Westerners Self-Enhance More than East Asians? *European Journal of Personality*, 23, pp. 183 - 203. doi: 10.1002/per.715

Ferguson, C.J. (2016, March 30). The Reduction of Ego-Depletion. *Huffington Post*, retrieved from http://www.huffingtonpost.com/christopher-j-ferguson/the-reduction-of-egodeple_b_9554874.html

Friese, M. & Fiedler, K. (2010). Being on the lookout for validity: comment on Sriram and Greenwald (2009). *Experimental Psychology*, 57-3, pp. 228 - 232. doi: 10.1027/1618-3169/a000051

Fiedler, K., Messner, C. & Bluemke, M. (2006). Unresolved Problems with the "I", the "A" and the "T": A logical and psychometric critique of the Implicit Association Test. *European Review of Social Psychology*, 17, pp. 74 - 147. doi: 10.1080/10463280600681248

Fiedler, K. & Hütter, M. (2012). The limits of automaticity. In Sherman, J., Gawronski, B. & Trope, Y. (Eds.), *Dual Processes in Social Psychology* (pp. 497-513). New York: Guilford Publications, Inc

Forscher, P.S., Lai, C.K., Axt, J.R., Ebersole, C.R., Herman, M., Devine, P.G. & Nosek, B.A.(2016). A Meta-Analysis of Change in Implicit Bias. (Manuscript under review). Retrieved from https://osf.io/awz2p/

Gaertner, S.L. & McLaughlin, J.P. (1983). Racial Stereotypes: Associations and Ascriptions of Positive and Negative Characteristics. *Social Psychology Quarterly*, 46-1, pp. 23 - 30.

Gawronski, B., Morrison, M., Phills, C.E & Galdi, S. (2017). Temporal Stability of Implicit and Explicit Measures: A Longitudinal Analysis. *Personality and Social Psychology Bulletin*, 43-3, pp. 300 - 312. doi: 10.1177/0146167216684131

Greenwald, A.G. (n.d.). Implicit Association Test: Validity Debates. *Washington University Faculty*, retrieved from http://faculty.washington.edu/agg/iat_validity.htm

Greenwald, A.G. & Banaji, M.R. (1995). Implicit Social Cognition: Attitudes, Self-Esteem and Stereotypes. *Psychological Review,* 102-1, pp. 4 - 27.

Greenwald, A.G., Banaji, M.R., Rudman, L.A., Farnham, S.D., Nosek, B.A. & Mellott, D.S. (2002). A Unified Theory of Implicit Attitudes, Stereotypes, Self-Esteem and Self-Concept. *Psychological Review*, 109-1, pp. 3 - 25. doi: 10.1037//0033-295X.109.1.3

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, pp. 553 -561. http://dx.doi.org/10.1037/pspa0000016

Greenwald, A.G. & Krieger, L.H. (2006). Implicit Bias: Scientific Foundations. *California Law Review*, 94-4, p. 945 - 967. Retrieved from http://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=1250&context=californialawreview

Greenwald, A.G., McGhee, D.E. & Schwartz, J.L.K. (1998). Measuring Individual Differences in Implicit Social Cognition: The Implicit Association Test. *Journal of Personality and Social Psychology*, 74-6, pp. 1464-1480.

Greenwald, A.G., Nosek, B.A. & Banaji, M.R. (2003). Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm. *Journal of Personality and Social Psychology*, 85-2, pp. 197 - 216. doi: 10.1037/0022-3514.85.2.197

Greenwald, A.G., Nosek, B.A., Banaji, M.R. & Klauer, K.C. (2005). Validity of the Salience Asymmetry Interpretation of the Implicit Association Test: Comment on Rothermund and Wentura (2004). In Bargh, J.A. (Ed.), *Automatic processes in social thinking and behavior* (pp. 265 - 292). New York, NY: Psychology Press. Retrieved from: https://faculty.washington.edu/agg/pdf/Nosek%20&%20al.IATatage7.2007.pdf

Greenwald, A.G., Poehlmann, 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, pp. 17 - 41. doi: 10.1037/a0015575

Greenwald, A.G. & Sriram, N. (2010). No Measure is Perfect, But Some Can be Quite Useful: Response to Two Comments on the Brief Implicit Association Test. *Experimental Psychology*, 57-3, pp. 238 - 242. doi: 10.1027/1618-3169/a000075

Hagger, M.S., Chatzisarantis, N.L.D., Alberts, H., Anggono, C.O., Batailler, C., Birt, A.R., . . . & Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science,* 11-4, pp. 546 - 573.

"Implicit Bias". (n.d.). Retrieved February 28, 2017 from Perception Institute website: https://perception.org/research/implicit-bias/

"Implicit Bias Resources". (n.d.). Retrieved March 2, 2017 from UCLA website: https://equity.ucla.edu/programs-resources/educational-materials/implicit-bias-resources/

Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, 2-8, e124.https://doi.org/10.1371/journal.pmed.0020124

John, L.K., Loewenstein, G. & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices with Incentives for Truth Telling. *Psychological Science*, 23-5, pp. 524 - 532. doi: 10.1177/0956797611430953

Kahneman, D. (2011). *Thinking Fast and Slow*. New York, NY: Farrar, Straus and Giroux

Lakatos, I. (1970). Falsification and Methodology of Scientific Research Programmes. In Kourany, J.A., (Ed.), *The Validation of Scientific Knowledge* (pp. 170 - 196). Boston, MA: Cengage Learning, Inc.

Lane, K.A., Banaji, M.R., Nosek, B.A. & Greenwald, A.G. (2007). Understanding and Using the Implicit Association Test: IV. What We Know (So Far) about the Method. In Wittenbrink, B. & Schwarz, N. (Eds.), *Implicit Measures of Attitudes* (pp. 59 -102). New York, NY: Guilford Press

Levy, N. (2012). Are you racist? You may be without even knowing it. *The Conversation*, retrieved from http://theconversation.com/are-you-racist-you-may-be-without-even-knowing-it-10826

Machery, E. (2016). De-Freuding Implicit Attitudes. In Brownstein, M. & Saul, J. (Eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology* (pp. 104 - 129). New York, NY: Oxford University Press.

McConnell, A.R. & Leibold, J.M. (2001). Relations among the Implicit Association Test, Discriminatory Behavior, and Explicit Measures of Racial Attitudes. *Journal of Experimental Social Psychology*, 37, pp. 435 - 442. doi: 10.1006/jesp.2000.1470

Mierke, J. & Klauer, K.C. (2001). Implicit Association Measurement with the IAT: Evidence for Effects of Executive Control Processes. *Zeitschrift für Experimentelle Psychologie*, 48-2, pp. 107 - 122. Retrieved from: https://faculty.washington.edu/agg/IATmaterials/PDFs/Mierke_Klauer_zexpsy_48_2001.OCR.pdf

Mooney, C. (2014, December 8). Across America, whites are biased and they don't even know it. *Washington Post*, retrieved from https://www.washingtonpost.com/news/wonk/wp/2014/12/08/across-america-whites-are-biased-and-they-dont-even-know-it/?utm_term=.d9c3ef19c351.

Mooney, C. & Viskontas, I. (2014, May 9). The Science of Your Racist Brain. *Mother Jones*, retrieved from http://www.motherjones.com/environment/2014/05/inquiring-minds-david-amodio- your-brain-on-racism/

Musgrave, A. & Pigden, C. (2016). Imre Lakatos. *Stanford Encyclopedia of Philosophy*, retrieved from https://plato.stanford.edu/entries/lakatos/

Ngwetsheni, S. (2016). Hair, Implicit bias and how cultural intelligence can heal our schools. *LeadSA*, retrieved from http://www.leadsa.co.za/articles/193709/hair-implicit-bias-and-how-cultural-intelligence-can-heal-our-schools

Nosek, B.A., Spies, J.R. & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth over Publishability. *Perspectives on Psychological Science*, 7-6, p. 615 - 631. doi: 10.1177/1745691612459058

Nosek, B.A., Greenwald, A.G. & Banaji, M.R. (2005). Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity. *Personality and Social Psychological Bulletin*, 31-2, pp. 166 - 180. doi: 10.1177/0146167204271418

Nosek, B.A., Greenwald, A.G. & Banaji, M.R. (2007). The Implicit Association Test at Age 7: A Methodological and Conceptual Review. In Bargh, J.A., (Ed.), *Automatic processes in social thinking and behavior* (pp. 265 - 292). Hove, United Kingdom: Psychology Press Retrieved from https://faculty.washington.edu/agg/pdf/Nosek%20&%20al. IATatage7.2007.pdf

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349-6251, pp. 943 - 952. doi: 10.1126/science.aac4716

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, pp. 171 - 192. doi: 10.1037/a0032734

Picket, K. (2017, March 27). Military Chaplains Slam Obama Admin Directive On 'Implicit Bias' Training. *The Daily Caller,* retrieved from http://dailycaller.com/2017/03/27/military-chaplains-slam-obama-admin-directive-on-implicit-bias-training/

Rachlinski, J.J. & Parks, G.S. (2009). Barack Obama, Implicit Bias and the 2008 Election. *Cornell Law Faculty Publications*, 1085. Retrieved from: http://scholarship.law.cornell.edu/facpub/1085

"Racism" (n.d.). *Oxford Living Dictionary*, retrieved at September 7 2017 from https://en.oxforddictionaries.com/definition/racism

"Reconsidering Implicit Bias" (2017, January 12). *Daily Nous*, retrieved from http://dailynous.com/ 2017/01/12/reconsidering-implicit-bias/

Rothermund, K. & Wentura, D. (2004). Underlying Processes in the Implicit Association Test: Dissociating Salience from Associations. *Journal of Experimental Psychology: General,* 133-2, pp. 139 - 165. doi: 10.1037/0096-3445.133.2.139

Schwarz, J. (1998, September 29). Roots of unconscious prejudice affect 90 to 95% of people, psychologists demonstrate at press conference. *UWNews*, retrieved from http://www.washington.edu/news/1998/09/29/roots-of-unconscious-prejudice-affect-90-to-95-percent-of-people-psychologists-demonstrate-at-press-conference/

Simmons, J.P., Nelson, L.D. & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data-Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science,* 22-11, pp. 1359 - 1366. doi: 10.1177/0956797611417632

Singal, J. (2017, January 11). Psychology's Favorite Tool for Measuring Racism isn't Up to the Job. *New York Magazine*, retrieved from http://nymag.com/scienceofus/2017/01/psychologys-racism-measuring-tool-isnt-up-to-the-job.html.

Stroebe, W. & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science,* 9-1, pp. 59-71. doi: 10.1177/1745691613514450

Teige-Mocigemba, S., Klauer, K.C. & Sherman, J.W. (2010). A practical guide to Implicit Association Tests and related tasks. In Gawronski, B. & Payne, B.K. (Ed.), *Handbook of*

implicit social cognition: Measurement, theory and applications (pp. 117 - 139). New York, NY: Guilford Press.

Tetlock, P.E. & Mitchell, G. (2008). Callibrating Prejudice in Milliseconds. *Social Psychological Quarterly*, 71-1, pp. 12 - 16.

The Washington Times (2016, October 3). Hillary Clinton's 'implicit bias'. Retrieved from http://www.washingtontimes.com/news/2016/oct/3/hillary-clintons-implicit-bias/

Van Tuijl, L., Glashouwer, K.A., Bockting, C.L.H., Tendeiro, J., Penninx, B.W.J.H. & De Jong, P.J. (2016). Implicit and Explicit Self-Esteem in Current, Remitted, Recovered and Comorbid Depression and Anxiety Disorders: The NESDA Study. *PLOS One,* 11-11, e0166116. https://doi.org/10.1371/journal.pone.0166116

Volkskrant (2016). Ben jij je bewust van je vooroordelen? Retrieved from https://www.volkskrant.nl/kijkverder/2016/vooroordelen/

Weber, C. (2016, November 16). University Holds Implicit Bias Training for Department of Public Safety and CNY Law Enforcement. *Syracuse University News*, retrieved from https://news.syr.edu/2016/11/university-holds-implicit-bias-training-for-department-of-public-safety-and-cny-law-enforcement-38060/

7. Acknowledgements

I hereby thank Diederik de Ceuster for his critical reviews during several stages of this project, and Maarten van Doorn for providing a helpful final review and discussion. I also wish to thank Jan Eisinga, Benjamin van Middelkoop, Joris van den Berg, Steffan Widdershoven and Simone Zwitserloot for their strong critiques and questions at an early stage of this thesis. Lastly, I wish to thank Brigitte Hooijmans, Ewald Oude Maatman and Deniz Sevim for their support and patience.