Neural correlates of the motivation to be moral

Nunspeet, F. van

Neural correlates of

the motivation to be moral

Félice van Nunspeet

The research reported in this dissertation was supported by the SNS-Reaal KNAW-Merian prize and the NWO-Spinoza prize awarded to Naomi Ellemers by the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO) respectively.

Printed by: Ridderprint B.V., Ridderkerk, The Netherlands (2014)

ISBN: 978-90-5335-863-4

Neural correlates of

the motivation to be moral

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens het besluit van de College voor Promoties
te verdedigen op dinsdag 27 mei 2014
klokke 11:15 uur

door

Félice van Nunspeet

geboren op 29 april 1986
te 's Gravenhage

Promotiecommissie:

Promotor:

Prof. Dr. N. Ellemers

Co-promotores:

Dr. B. Derks

Overige leden:

Prof. Dr. E. A. M. Crone

Prof. Dr. S. T. Nieuwenhuis

Prof. Dr. K. van den Bos Universiteit Utrecht

Prof. Dr. D. H. J. Wigboldus Radboud Universiteit Nijmegen

Contents

Chapter 1

General Introduction

This dissertation addresses a well-known but vast topic: Morality. Previous research has revealed that it is important for people to be moral. Nevertheless, they may sometimes commit immoral acts. In this dissertation, I take a social psychological perspective from which I examine when and why people become motivated to do what is right. I study whether people tend to adhere to their own moral values, and whether their moral behavior is affected by the presence of others. Moreover, by borrowing research methods from neuroscience, I aim to unravel some of the brain processes involved in this motivation to be moral.

Previous Research on Morality

Researchers across scientific disciplines, who examine different aspects of morality, work on the assumption that people have an innate sense of what is right and wrong. In fact, some of these researchers even argue that moral behavior is not unique for humans, but reflects a more basic concern for the well-being of others, that we share with some animals. For example, De Waal studied aspects of morality in chimpanzees, bonobos and capuchin monkeys. Results of his studies revealed that such animals show fairness concerns: When precious goods – such as attractive food items – are not equally distributed, they show signs of resentment (Brosnan & De Waal, 2003). Moreover, they comfort each other in distress and cooperate with other individuals in need of help, even if there is no immediate gain for the self (De Waal & Berger, 2000; see also De Waal, 1996). The fact that such indications of cooperation and empathy are found in primates (as well as other animals, such as elephants) is often interpreted as evidence that moral behavior represents a very basic and almost instinctive tendency – also for humans.

In the study of human behavior, developmental psychologists have theorized about how morality is established in childhood and develops through adolescence and adulthood (e.g., Kohlberg, 1969; Piaget, 1965). More recently, neuroscientific researchers have examined the effects of damage to (prefrontal) parts of the brain and have shown that such impairments are associated with immoral conduct and unethical decision making (e.g., Anderson, Bechara, Damasio, Tranel, & Damasio, 1999; for a review see also Moll, Zahn, De Oliveira-Souza, Krueger, & Grafman, 2005). These approaches thus also suggest that people have an intrinsic sense of morality (or a so-called moral intuition; see for example Haidt, 2001). Variations in

moral behavior seem to stem only from differences in the extent to which morality is developed in childhood or impaired due to physical restraints in the brain.

One could thus argue that people do not need explicit guidelines for what is the right thing to do, as they know this intuitively. This resonates with the consensus among researchers that moral principles are universal and fundamental to who we are. Yet, we are confronted with people's immoral acts on a daily basis: Every news website and –paper contains examples of people lying, stealing and cheating. Knowledge of the person who committed such an immoral act may surprise us. The people who are known for their good intentions, can still decide to act immorally. Likewise, research shows that the same individuals may show moral as well as immoral behaviors at different points in time (e.g., Monin & Miller, 2001). Why is this the case?

A Social Psychological Perspective on Morality

Prior attempts to answer this question have mainly investigated why people transgress moral norms. In line with the assumption that moral behavior is a natural tendency, such transgressions can be attributed to deficiencies in personal moral development or to cognitive limitations preventing people from showing 'regular' moral behavior. In this dissertation, I take a social psychological approach. I work on the notion that it is 'normal' for individuals to shift their moral behavior across situations or over time. I explicitly study these variations, focusing on *situational* features that induce moral behavior, as a starting point to increase our understanding of why and when people *adhere* to moral norms. Thus, the central aim of my research is to uncover which social mechanisms enable people to behave in line with their (and other people's) moral values, and how this affects the way they approach different situations. I argue that, by using this approach, we will gain a better understanding of how moral behavior can be stimulated by situational features. This can help bring out the best in people, regardless of their individual differences. To achieve this goal, I address three questions in the current dissertation: (1) Do people tend to act in ways that are considered moral? (2) How important is it for them to be perceived as moral by others? (3) How much do they care whether or not they succeed in behaving according to their moral values?

Examining these questions from a social psychological perspective means that I take into account the impact of how people see themselves, how they are judged by others, and to which social group they belong. In addition to examining social psychological factors in explaining displays of moral behavior, I use neuroscientific and psychophysiological indicators that may reveal the cognitive and affective mechanisms underlying such behavior. Combining these different approaches makes it possible to go beyond self-reported statements about what people say they will do. This also allows me to examine any discrepancies between the way people actually behave, and what they explicitly report. Going beyond prior work, my aim is to reveal whether and how people act upon their moral values by examining cognitive processes associated with moral behavior.

Diverging Perspectives on Morality

To examine what motivates people to be moral, we first must know what "being moral" actually means. In books of law or religion, morality is often defined by specifying what is *not* moral. The origin of current notions about human rights and general behavioral guidelines ('though shall not steal') can thus be traced throughout history and converges across national contexts, cultures, and religions. When moral standards are not made explicit, we may however still be guided by our moral intuition: An undefinable but certain intuitive state that indicates that something is right or wrong (Haidt, 2001).

The central goal of moral behavior thus seems quite obvious: Doing what is right. However, how this takes form in a concrete manner or in a specific situation is much more ambiguous. You may have noticed that what you consider the right thing to do may differ, depending on particular circumstances, or the presence of other people. For example, you know that helping others is generally considered moral. Nevertheless, you may be more motivated to help your friends or family members than some stranger in the street. In a similar vein, you are likely to care whether others perceive you as a moral person. At the same time, opinions of others you consider relevant to yourself – such as your friends or family – are likely to matter more than opinions of people you do not know. As a consequence, whether or not you act upon general moral guidelines is likely to differ, depending

on who is affected by your behavior, or who are present to observe and evaluate your behavior.

Individuals deliberate about what would be the right thing to do, but so do groups, institutions and countries. To give an example, let us consider the Olympic Winter Games of 2014. Ever since the Olympic Committee announced that these Games would be hosted by Sochi –Russia, one could hear objections around the world. Several countries objected to the organization of an event that promotes peace and international cooperation through sports, by a country that is associated with limited civil rights such as rights of freedom of assembly and freedom of speech. Also in the Netherlands, there were fierce discussions about the decision of the government to send a large political delegation (in addition to members of the royal family) to attend the Games. According to protesters, this signaled the wrong message: Given the high moral standards concerning civil rights in the Netherlands, this country should not support an event organized and propagated by another country that violates such rights. Such debates thus raise questions about national moral values, how to (re)present those values, and how these will be perceived by other communities and countries.

The above example illustrates that there can be differences between groups in moral values on an international level. Debates about what is right may however also divide different groups within the same country. For example, in the Netherlands, Belgium, and France, discussions concerning the integration of Muslims invoke moral concerns. Norms posed by the Islam seem to oppose common societal practices in Western countries. This is the case for instance with the clothing habits that many Muslims endorse, such as wearing a burka or a hijab – a headscarf– for Muslim women. Wearing a burka in public was prohibited in France in 2011. A similar judicial proposition was discussed in the Netherlands as well. Wearing such clothing may be perceived in Western societies as degrading for women and as morally wrong because it could strengthen the segregation of Muslim and non-Muslim individuals. In contrast, Muslims see this as a sign of modesty and high moral standards. This illustrates that the same behavior (such as wearing a headscarf) can be considered the moral thing to do by some, while being

seen as immoral by others. In other words, it is not always easy to specify the 'right' thing to do because each group may have its own moral norms.

In a context where members of multiple groups are present, people can therefore question what would be the moral thing to do. When the former queen of the Netherlands went to Oman for a state visit, she wore a headscarf whenever she visited a mosque. She argued that she did this out of respect and regard for the country, its people and their religion. Several members of the Dutch government supported her judgment. However, there were also politicians who openly condemned her opinion and related behavior. This example thus also illustrates that debates about what is moral touch upon who we are as individuals, and how we see ourselves in relation to our groups (e.g., a political party, 'the Dutch'). They also concern our moral principles and values; how we want to portray ourselves to others; and how we want to be perceived by them. These are questions that are central to the current dissertation.

Morality and Group Inclusion

The importance of the people around us for how we think about ourselves and decide upon how to behave can be explained from a social identity approach. Social identity theory posits that people often perceive themselves and others as part of a group. Groups help people to define who they are, where they belong, and how they should behave (Tajfel & Turner, 1979). Being part of a group with whom one can share his or her social identity (e.g., "the Dutch", "social psychologists") is a way to validate one's self-views, and to establish and maintain one's self-esteem (see also Ellemers & Jetten, 2013). Groups thus can help people to establish a distinct identity: Groups each have their own norms which make them different from other groups. The norms and values within a group thus provide clear guidelines as to how individuals should behave in order to secure inclusion in that group. As a result, people tend to look for inclusion in a group with whom they can share their moral values and principles. Alternatively, they adapt their own values to the groups that are important to them. It thus depends on whether people want to belong to and identify themselves with a particular group (whether they consider this their 'ingroup') whether they adjust their behavior according to the groups' moral norms. This refines the idea that people's

behaviors are affected by 'social pressure' in general. People care primarily about adherence to norms within their ingroup, while it is less important for them to behave according to outgroup standards. For example, Dutch Muslim women who identify more strongly with their religious group than their nationality are more likely to adhere to the norms of their religious group (e.g., by wearing a headscarf) than the norms of their Dutch nationality (e.g., not wearing a headscarf).

Group norms and standards are particularly important when these relate to morality. As a member of a group, people are more inclined to adhere to ingroup norms when these are presented as "the moral thing to do", rather than prescribing what would be "the competent thing to do" (Ellemers, Pagliaro, Barreto, & Leach, 2008). People do this because they think they will receive respect from their fellow group members when they adhere to moral group norms (Pagliaro, Ellemers, & Barreto, 2011). Moreover, people identify more strongly with a moral than a competent group and are more proud to be member of groups that can contribute to their morality than groups that stand out for their competence (Leach, Ellemers, & Barreto, 2007).

Morality also seems to be the most important determinant of the impression we form of other individuals and groups. When encountering someone we do not know, we primarily search for characteristics indicating their morality (e.g., honesty, trustworthiness) rather than showing an interest in competence (e.g., particular skills, intelligence) or sociability (e.g., kindness, friendliness; Brambilla, Sacchi, Rusconi, Cherubini, & Yzerbyt, 2012).

Thus, both at an individual and a group level, people look for characteristics concerning morality –rather than information concerning other people's competence or sociability (Brambilla, Rusconi, Sacchi, & Cherubini, 2011). In fact, research shows that we are able to determine whether another person is trustworthy in less than a second. This happens even faster than making judgments about whether that person is attractive, competent, or nice (Willis & Todorov, 2006). In the process of gathering information about how moral someone is, special importance is attached to any negative behaviors. That is, we more likely to conclude that someone is immoral when s/he has done something wrong, than we are to conclude that this person is moral because s/he is always honest and reliable.

In other words, even for a person who is known for his or her moral integrity, a single act of immoral conduct can spoil this positive image, because immoral acts are perceived as more informative of someone's true character than moral acts (Skowronski & Carlston, 1987).

This is not only important when learning about someone else's moral characteristics and values, but also plays a role in the concerns people have about *themselves* being seen as moral by others. That is, if one's morality is called into question, then one's identity and sense of self is negatively affected. For example, when there is disagreement about moral values (as compared to material interests), or when a person is evaluated on his or her prior immoral behavior, people report increased negative affect and display a physiological threat response (Kouzakova, Harinck, Ellemers, & Scheepers, 2014; Van der Lee, 2013).

When others question their moral intentions or behaviors, people worry that they may lose respect or even will be excluded from the group. However, since the meaning of morality differs between groups and situations, it can be impossible to do what is right according to everyone. People may therefore focus primarily on doing what is right according to their own ingroup. Such ingroup norms may however also concern how one should behave towards members of another group (e.g., treating people from other cultures with respect). The intention to adhere to such ingroup norms may be relatively easy as long as interactions with outgroup members are hypothetical. But what happens when people are faced with an actual interaction with a member of another group? For example, when non-Muslim have to collaborate with a Muslim at work?

Morality in Intergroup Relations

As I explained above, morality plays an essential role in regulating individual behavior within a person's own group. It is however just as important in intergroup interactions. Accordingly, morality is often examined in such contexts. For example, Reed and Acquino (2003) revealed how intergroup conflict can be diminished by extending ingroup favoritism towards individuals representing different religions and ethnicities. That is, people show increased explicit moral regard towards outgroup members when they attribute greater importance to their moral identity (Acquino & Reed, 2003). Likewise, previous research has shown that

people's willingness to strive towards social equality between groups is enhanced when other ingroup members say this is an important moral ideal (rather than when they say it is a moral obligation; Does, Derks, & Ellemers, 2011). Evaluating or presenting people's identity or behavior in terms of moral values can thus enhance their moral intentions and acts. In other words, telling people that they should act according to what they think is the right thing to do may thus be used as an instrument to enhance moral behavior towards and between people. I assess the implications of the effects of such an emphasis on a person's morality, in the current dissertation. Specifically, I examine what happens when the implications of behavior of native Dutch (non-Muslim) individuals towards Muslim women are presented as an indication of their egalitarian values. I propose that reminding people that their behavior conveys their morality will stimulate equal treatment and motivate people to avoid displaying bias towards the Muslim outgroup.

Measuring Moral Behavior

Thus far I have discussed why it may be important for people to adhere to their own moral values and the moral norms within their groups. If people indeed want to be perceived as moral, this could cause them to emphasize the importance they attach to moral behavior because they think this may reflect positively upon the image others have of them. This may however not necessarily reflect their actual behavioral preferences. Nor does it predict how they would act in a specific situation, for instance when they do not realize that others are paying attention to their moral tendencies. In other words, people may deliberately respond in a socially desirable fashion when they think their moral image is at stake. This is a relatively common concern when interpreting responses to self-report questionnaires. Emphasizing the implications of people's behavior in terms of how moral they are could thus introduce measurement problems. Relying on self-reported intentions to assess people's responses may not capture their 'true' intentions, or their intentions may not correspond to their actual behavior. In this dissertation, I therefore used another type of measure to assess the motivation to be moral. I adapted an Implicit Association Test (IAT) to examine participants' behavioral responses that might reveal bias favoring their own ingroup (non-Muslims) over members of a relevant outgroup (Muslims).

The IAT was first developed by Greenwald, McGhee, and Schwartz (1998) to assess the strength of (automatic) associations between target concepts and different attributes. This test assumes that people find it easier to quickly connect concepts that they implicitly relate to each other. You can imagine how this works when you are asked to couple a concept, such as "flowers", with words like "fun" or "kind" (i.e., attributes). Making such connections should be relatively easy because "flowers", as well as words like "fun", both have a positive connotation in our mind We thus associate one with the other, because they are what is called *congruent*. Likewise, it should be relatively simple to couple a concept such as "bugs" with words like "pain" or "fear". In this case, the association is easily made because both the concept and the words have a negative connotation. However, things are likely to become more difficult when you try to couple "flowers" with "pain", or "bugs" with "fun". This is because these concepts and words do not have the same connotation - a positive word has to be coupled with a negative concept - and are thus *incongruent*. They are therefore not easily associated with one another.

An IAT is based on these associative mechanisms. It is a reaction time task during which participants are asked to press one key as quickly as possible when they see a particular word or picture. In one part of the task, they are asked to respond with the same key to both pictures or names of "flowers" and positive words (e.g., "fun"). They are asked to press another key for both pictures and names of "bugs" and negative words (e.g., "pain"). This procedure is used to assess participants' performance on congruent trials. In another part of the task, the pairing becomes less intuitive. Here, participants are asked to respond with the same key to both "flowers" and negative words. Another key has to be used to indicate both "bugs" and positive words. These instructions are used to assess participants' performance on incongruent trials. To the extent that people are more inclined to associate flowers with positivity and bugs with negativity, they should respond more quickly to congruent trials than incongruent trials. Thus, the difference in their reaction times on incongruent compared to congruent trials reveals the strength of their implicit associations. This is what is called the IAT effect. It indicates the extent to which people find it more difficult to associate one concept (e.g., "flowers) with negative rather than positive words, and another

concept (e.g., "bugs") with positive rather than negative words. The difficulty of making such associations is revealed in increased response times. In this way, the IAT effect can reveal people's negative bias towards all kinds of manner of target concepts, including bugs.

The example of flowers and bugs illustrates the principles on which the IAT effect is based. However, the test has most often been used to assess implicit negative bias towards different groups of people in society, in studies concerning prejudice. In this case, you are also asked to couple a concept with positive words. But this time, the concept is not "flowers", but represents a social group, for instance native Dutch people. As indicated above, people are concerned with having a positive social identity. Hence, they are likely to think more positively of groups associated with the self (ingroups) than of other groups (outgroups). The groups to which they belong (and Dutch participants can be seen to belong to the group "the Dutch") is likely to have a positive connotation. In comparison, people are more likely to have negative connotations with an outgroup, such as immigrants. When performing the IAT, responding with one key to the concept "native Dutch" and positive words may thus be relatively easy, as is responding with another key to the concept "immigrants" as well as negative words, as these represent congruent associations. In contrast, responding with a single key to the concept "native Dutch" and negative words is likely to be more difficult –and will therefore take more time– just as responding with another key to the concept "immigrants" as well as positive words (incongruent associations). The IAT effect (i.e., the difference in response times between incongruent and congruent trials) in this case reveals the extent to which it is more difficult to associate one's ingroup (e.g., "native Dutch") with negative rather than positive words, and an outgroup (e.g., "immigrants") with positive rather than negative words. In other words, the IAT score can reveal people's implicit negative bias (prejudice) towards immigrants.

The target concepts "native Dutch" and "immigrants" are an example of concepts that can be used in an IAT. In the United States, the IAT is often used within a racial context. Such a 'race IAT' consists of stimuli (such as photographs) representing people with a white or dark toned skin color. Explanations for white people's tendency to reveal a negative bias towards black people are diverse. People

with a dark skin tone may be seen as outgroup members by people with a white skin tone. The differentiation between these two groups may thus reveal positive associations with the ingroup and negative associations with the outgroup. However, in the case of the 'race IAT' other explanations could also be offered for this pattern of associations. For example, stereotypes of people with a dark skin tone may more often be negative rather than positive. Think for example about stereotypes concerning criminal records and aggression. As a result, the physical features of a black man's face may be perceived as more threatening than the physical features of a white man's face, which could cause negative rather than positive associations with this type of stimulus. All these explanations could thus explain the emergence of negative bias on the IAT, against people with a dark skin tone.

In the current dissertation I use different target concepts in the IAT because of two reasons. First, negative stereotypes concerning people with a dark skin color as well as discrimination against this group are less common in the Netherlands. Such a 'race IAT' is thus less relevant to assess in a Dutch research population. Second, I attempt to rule out some of the additional explanations for a negative outgroup bias –besides the explanation of one's social and distinct social identity. The IAT target concepts I use in this dissertation are "women without a headscarf" and "women with a headscarf". Women without a headscarf represent native Dutch, non-Muslim women. These women are similar to my research participants and thus are likely to be seen as ingroup members. Women with a headscarf represent Muslim individuals. These women are different from my research participants and thus are likely to be perceived as outgroup members. As I indicated in the first part of this introduction, the integration of Muslims is a current topic of debate in the Netherlands. This debate to an important extent addresses clothing habits, such as wearing a headscarf, in public places or functions. Measuring people's negative bias against Muslim women (i.e., women who wear a headscarf) is thus more relevant for research in the Netherlands. Furthermore, I pretested the photographs of the faces of these women (i.e., the stimuli in the IAT) on different characteristics. Examples are perceived kindness, honesty, intelligence, and attractiveness. Results of this pretest showed that both the women with and

without a headscarf are perceived as equal concerning these characteristics (see Appendix A of this dissertation for more details). A negative bias against women with a headscarf –as revealed by this IAT– can thus not be explained by any negative associations (related to such characteristics) with the stimulus materials as such. Importantly, the Muslim women presented in the IAT were only perceived as *different* from the research participants –but not in any way more negatively or less positively than the non-Muslim women. If I find a difference between positive and negative associations with Muslim and non-Muslim women this can therefore only be attributed to the fact that the women with a headscarf are being perceived as different from the research participants – i.e. as outgroup members. In other words, the use of these stimulus materials implies that any negative bias against Muslim women that is revealed by the IAT, can only be attributed to the fact that these individuals are seen as representing another (out)group.

Emphasizing the Implications of One's Behavior

In my research, I use the IAT as an indicator of people's negative bias, or prejudice, towards outgroups. Some would propose that the associations between groups and positive and negative attributes that people make during the IAT, are made easily and quickly because they occur automatically. However, prior research has revealed that IAT performance is malleable: The fast elicited response to associate some concepts and attributes are not automatic, but can be adapted. That is, participants can deliberately influence their performance by using strategies that diminish the difference between response patterns on congruent and incongruent trials. This is the case, for instance, when they are informed about how their bias will be measured (Fiedler & Bluemke, 2005). Likewise, IAT responses are adapted when people are explicitly motivated to enhance their self-image or to emphasize their positive relationship with other individuals (for an overview see Blair, 2002). Thus, using an IAT, it is possible to examine whether people adjust their performance when motivated to do so. In the current dissertation I examined whether participants performed differently when they were reminded of the moral implications of their behavior during this task. That is, I examined whether people showed less implicit bias against Muslim women when they thought that the test would reveal whether they treated Muslims and non-Muslims equally (e.g., their

moral values concerning egalitarianism and discrimination), rather than merely being good at quickly processing information and learning to make new associations.

Because of the stimuli used in an IAT, performance on the test can relatively easy be seen as indicating prejudice, and thus perceived as a measure of moral values. However, at the same time, the test is a reaction time task in which participants are asked to sort different types of stimuli according to changing rules. The faster and more accurately participants respond, the better their performance. Thus, it would be equally plausible to see the IAT as a test of people's ability to perform well on this task. In other words, the IAT can be perceived both as a measure to detect social bias against an outgroup, *and* as a measure of one's competence. My aim is to examine whether people respond differently during the IAT depending on which of these task implications is emphasized. This allows me to investigate whether (and how) people adjust their behavior when they think their performance can indicate their moral values concerning egalitarianism.

In most of the studies reported in this dissertation (i.e., Chapters 2 through 5), the IAT is used to assess behavioral responses. This approach extends previous research concerning the importance of morality for people's self-views and social identity, which has mainly relied on explicit self-report measures. Since people may adjust their deliberate responses on a self-report questionnaire to convey what they think is perceived as moral by others, their answers may not necessarily reflect the way they will actually respond in situations where moral concerns play a role. Assessing people's moral responses in a less explicit way, by using this IAT, can thus provide insight in whether people actually behave in line with relevant moral values. In addition to assessing task behavior to reveal implicit bias, I use psychophysiological and neuroscientific research methods to increase our understanding of the cognitive processes that underlie people's adherence to their moral values.

The Added Value of Cognitive Neuroscience

An important additional aim of the research reported in this dissertation is to examine the cognitive and neural mechanisms that underlie people's motivation to act according to what is moral and to be perceived as moral by others. Previously, I

explained how behavioral performance on an IAT can give us more information about a person's 'true' behaviors. In addition, I aim to uncover *how* people monitor and adapt their behavior to achieve adherence to moral values. Understanding the cognitive processes that help people to behave in a moral manner may expand our knowledge of the mechanisms needed to behave morally. It may reveal whether people *initiate* their behavior in a different way when they think this is indicative of their moral values (as compared to, for example, their competence). People may for instance pay more attention to other people's skin color when they have just been informed about discrimination rates. As a result, they may more quickly detect someone with a different ethnic background which will help them to act in an unprejudiced manner. On the other hand, the cognitive mechanisms may reveal increased vigilance to errors, which may help people to *adjust or redirect* their responses to avoid displaying signs of possible immoral behaviors when they want to appear moral.

This type of behavioral initiation or correction is likely to occur outside of one's conscious awareness. In a job interview for instance, an employer may be focused on the applicant's gender because he read a report the day before about the under-representation of women in business organizations. At the same time, the employer may not be aware of his increased attention to that aspect of the applicant. He may not even consciously remember reading that report. When asked to verbalize his considerations, the employer may thus be unable to report that he was more focused to the applicants' gender. In fact, even when the employer was aware that he was more attentive to the gender of the applicants that day, he may not want to disclose this for fear of revealing gender bias. In other words, people may not be *able* to tell us about the cognitive processes that they recruited in order to behave in a particular way. And even if they are able, they may not be *willing* to tell us about those processes.

Neuroscientific research methods can help solve such problems. Methods used in cognitive neuroscience have proven to be effective in gaining insight in processes such as enhanced or decreased attention. Using such measures can thus reveal additional information about the mechanisms underlying people's actual behavior. They make it possible to capture automatic and/or unconscious response

tendencies elicited by moral situations. Additionally, such neuroscientific measures provide an unbiased perspective on what actually happens during task performance, as these indicators are not sensitive to people's supposedly heightened social desirability to comply to moral expectations.

Extending the Cognitive Neuroscience of Moral Reasoning

Cognitive neuroscientists have already begun to shed light upon the cognitive and neural mechanisms associated with moral reasoning and decision making. For instance, previous research has examined how people reason when they are confronted with a moral dilemma and asked to decide how they would behave in such a scenario. A famous example concerns the so-called 'trolley dilemma' in which people are asked whether they would sacrifice one person's life in order to save five other individuals (Foot, 1978; Thomson, 1985). Neuroscientific research has revealed that brain networks associated with both cognitive as well as emotional processes are involved in such moral reasoning (e.g., Greene, Nystrom, Engell, Darley, Cohen, 2004). Moreover, research concerning the judgment of moral and immoral acts has revealed that people are highly sensitive to the detection of moral transgressions which may be related to the instant emergence of moral emotions such as disgust (e.g., Borg, Lieberman, & Kiehl, 2008; Schnall, Benton, & Harvey, 2008). Such studies have thus focused on the mechanisms underlying people's individual ability to reason about and decide what is and what is not moral. However, as I have explained above, social contexts may affect what can be considered the moral thing to do. Likewise, different situations may affect whether people actually behave according to what is perceived as moral. These social factors are often neglected in cognitive neuroscience, as much of the research in this tradition focuses on establishing universal response patterns. Nevertheless, I argue that moral behavior is likely to shift across different contexts, depending on the social concerns that are raised. Additionally, knowing right from wrong and being able to make moral judgments may differ significantly from people's actual moral intentions, motivations, and subsequent behavior. Thus, to gain better understanding of people's motivation to adhere to their own moral values, and how they enact this motivation, I will investigate the mechanisms underlying actual

moral behavior (i.e., IAT performance), and how these are affected by different social contexts.

Multiple Research Methods

Besides using self-reports and measuring behavioral responses on the aforementioned IAT, I used three different research methods in the studies reported in this dissertation: Skin conductance, EEG, and fMRI.

Skin conductance.

Skin conductance indicates electrodermal activity representing activation in the sweat glands, measured at the skin surface of our hands. Skin conductance relates to so-called "psychologically induced sweating" (Dawson, Schell, & Filion, 2000, p. 202). People automatically sweat when they experience emotions, when they become aroused, or when their attention is increased. Measuring the tonic level of skin conductance can thus be used as an unobtrusive way to examine general states of arousal and alertness. Moreover, phasic skin conductance responses (SCRs) can be elicited by different characteristics of an occurring event. In psychological experiments this may be a particular stimulus that is new, intense, or has an emotional impact. Skin conductance is an automatic response generated by the sympathetic nervous system, a process that thus cannot easily be adapted by the participant for self-presentational reasons. Additionally, variations in skin conductance can be measured *while* participants receive relevant information. I am interested in whether people care about succeeding in behaving according to moral norms. Skin conductance is thus a valuable measure to detect how people (physically) respond to information indicating that they are, or are not, as moral as others.

EEG.

EEG is the abbreviation of *electroencephalogram*, which is an indicator of brain activation measured across the scalp (e.g., Luck, 2005). EEG has a relatively low spatial resolution: It is usually unclear from which brain region the activity originates, because it is measured at the scalp. This noninvasive neuroimaging technique does however have a high *temporal* resolution: Evoked responses in brain activation can be measured within milliseconds after a stimulus is presented on the screen or when a response is given.

An EEG can be used to monitor ongoing brain activation during a complete experiment. From this EEG, we can extract responses evoked by particular events –so-called event-related potentials (ERPs; Luck, 2005). Using ERPs, it is thus possible to gain insight in the (ongoing) cognitive processes associated with particular parts of the experiment. For example, ERPs around a given response can inform us about the preparation of and the reaction to that response on a cognitive level. This thus complements the actual behavioral response that can only indicate for instance what people decide to do, or how long it takes for a participant to make this decision.

Besides examining ERP's during task responses, we can also investigate how different stimuli are processed in the brain. In the IAT I developed for the research in this dissertation, the target concepts are presented by photographs. Specifically, the target concept "ingroup" is represented by photographs of women without a headscarf. The target concept "outgroup" is represented by photographs of women with a headscarf. Using ERPs, it is possible to detect that these two types of photographs are differentially processed in the brain. In addition, I can examine whether the ERP modulations associated with viewing ingroup and outgroup members are affected by people's motivation to perform in line with their moral values. In addition to a computation of the behavioral responses on the IAT (i.e., response latencies and the amount of accurate responses), ERPs can thus reveal the attentional processes associated with such task performance.

fMRI.

In contrast to EEG, functional Magnetic Resonance Imaging (fMRI) is a neuroimaging technique that has a relatively low temporal resolution but a high *spatial* resolution – it reveals which brain areas are activated (e.g., Huettel, Song, & McCarthy, 2004). Although also noninvasive, MRI is used to visualize internal physical tissue. Moreover, within the brain, *functional* MRI is used as an indicator of brain activation during task performance. Using the blood-oxygen-level-dependent (BOLD) response, differences in deoxygenated blood levels are measured. Performing a task elicits specific cognitive demands, such as increased attention. For such cognitive demands an increase in energy is needed in particular parts of the brain. One of those sources of energy is oxygen. Release of this oxygen

increases the level of deoxygenated blood which can be detected using magnetic resonance. This is thus used as the indicator of brain activation (Huettel et al., 2004). The whole brain can be visualized using MRI and the BOLD-response can be measured to localize activation in specific brain regions. It is thus possible to compare the degree and location of brain activation associated with different parts of an experimental task. This also implies that we can detect activation in *sub*cortical regions of the brain (that are located deep in the brain), including structures associated with primary affective responses. In addition to behavioral and self-report measures, but also extending information gathered with ERPs, fMRI can thus inform us about the brain regions involved in moral task performance.

Overview of the Dissertation

With the research reported in this dissertation, I address three different research questions. In Part I, I examine whether people tend to act in ways that are considered moral. in Part II, I investigate how important it is for people to be perceived as moral by others. Finally, in Part III, I focus on how much people care whether or not they succeed in behaving according to their moral values.

In Part I, I examine whether people tend to be more motivated to show that they are moral than that they are competent. To be able to make this comparison, I present an IAT as either indicative of one's moral values concerning egalitarianism and discrimination, or of one's ability to learn new tasks and to quickly process information. In Chapter 2, I examine whether participants' task performance is affected by this difference in emphasis on specific task implications. Specifically, I test the prediction that when the moral implications of the task are stressed participants show a weaker negative bias against Muslims than when the competence implications are emphasized. Additionally (using EEG in Chapter 2, and fMRI in Chapter 3), I examine whether the moral test implications enhance participants' attention towards different aspects of the task. In other words, these studies aim to reveal whether stressing the implications of one's behavior in terms of one's moral values causes people to adjust and direct the focus of their attention during task performance.

In Part II, I examine how important it is for people that others think they are moral. In this part of the dissertation, the implications of the IAT are again presented in terms of one's moral values or competence. Additionally, participants are led to believe that their performance on the IAT is being monitored and evaluated by someone else. In Chapter 4, I examine whether people show their motivation to be a moral group member by inhibiting their bias against Muslims when an ingroup rather than an outgroup member is evaluating their performance. In this chapter, the evaluator is a non-Muslim individual (who's gender is matched with that of the participant). In one condition, she is presented as someone with the same group membership as the participant. This is achieved with very minimal instructions (also referred to as the 'minimal group paradigm'; Tajfel, 1970). The participant is told that their evaluator has the same personality type as they do and that s/he is thus an ingroup member. In another condition, the evaluator is presented as someone representing the other minimal group. Participants in this condition are thus told that their evaluator has another personality type than they do and that s/he thus can be considered an outgroup member. As in Part I, I thus examine whether people adjust their behavior and increase their attention towards the task in case the moral implications are emphasized. In addition, I test whether participants are more inclined to do this when an ingroup member, rather than an outgroup member, is evaluating their behavior.

In Chapter 4, participants' IAT performance is thus monitored by a non-Muslim individual who is introduced as someone with the same or another personality type as the participant. Thus, the evaluator is introduced as someone who is similar to or different from the participant based on an implied personal feature. Nevertheless, the person evaluating participants is always the same man or woman, and the evaluator's visible appearance is always the same as that of the participant. In Chapter 5, I examine whether people's moral behavior towards an outgroup (i.e., appearing unprejudiced) is affected when they are being monitored by someone who can be seen to represent the target outgroup in the IAT: A woman with a headscarf. Such an evaluator can thus be seen as an outgroup member. In principle, being seen as moral by outgroup members should be less important than being seen as moral by ingroup members. This might imply that

participants should not be motivated to appear moral towards their Muslim evaluator – because she is an outgroup member. On the other hand, since the moral behavior assessed in this research is people's bias against Muslims, a Muslim evaluator (representing the target group against whom bias might be revealed) could still have an impact on participants IAT performance – albeit for different reasons (see also Lowery, Hardin, & Sinclair, 2001). I thus examine the effects of being evaluated by a Muslim woman on moral task performance. Additionally, I compare the impact of being monitored by this Muslim evaluator, depending on whether she is presented as an ingroup or an outgroup member based on the previously described minimal group membership. Specifically, participants are informed that their evaluator either has the same or another personality type as they do. I thus also examine whether presenting an outgroup member (the target group representative) as a partial ingroup on another dimension (same personality type) helps people to reduce prejudice against the target outgroup.

In Part III, I again address people's motivation to act according to what is considered moral. But here I go one step further. I focus on how much people care whether or not they *succeed* in behaving according to their moral values. In Part III, I aim to extend the findings of Part I, in which I examine whether and how people's motivation to be moral affects their performance on a measure of bias against Muslims. In Chapter 6, participants are provided with feedback about their performance on this test. They either are confronted with information indicating that they are less moral (or less competent) than other participants, or with information stating that they are more moral (or more competent) than other participants. I examine the emotional and psychological impact of this information. Specifically, I measure self-reported affect and skin conductance responses as an indicator of physiological arousal. If people care more about being moral than competent, receiving negative information about their own moral behavior should be more distressing than being confronted with negative information about one's competence. Nevertheless, or even because of this reason, people may be more attentive to positive information about their morality since this may confirm their moral self-concept. In an additional fMRI study I further examine this prediction

and compare whether information related to one's morality rather than one's competence is processed as more self-relevant in the brain.

In the final part of the dissertation (Chapter 7), I integrate the findings presented in the three empirical parts. I discuss their implications and how this research contributes to current insights in social psychology and social neuroscience, and I consider the societal implications of my findings. Note that Chapters 2, 3, 4, 5, and 6 are prepared as separate journal articles. This results in some overlap in the theoretical background and method sections, but also implies that these chapters can be read independently.

Part I

The importance of being moral

Chapter 2

Moral concerns affect implicit prejudice and associated cognitive processes: Behavioral and ERP findings

This chapter is based on:

Van Nunspeet, F., Ellemers, N., Derks, B., & Nieuwenhuis, S. (2014). Moral concerns increase attention and response monitoring during IAT performance: ERP evidence. *Social, Cognitive, and Affective Neuroscience, 9,* 141-149. doi:10.1093/scan/nss118

We tend to evaluate people's personal characteristics and behavior along two dimensions: One concerning morality (i.e., how we should behave) and one concerning competence (i.e., how we are able to behave). Behaving according to these dimensions is differentially diagnostic for who we are and how we are perceived: Skowronski and Carlston (1987) showed that for morality negative behaviors are perceived as more diagnostic than positive behaviors, whereas for competence positive behaviors are more diagnostic than negative behaviors. In contrast to behaving incompetently, behaving immorally thus seems to be more indicative of who we are.

Recent research has shown that for people's self-views and the positive evaluation of the group to which they belong moral characteristics are perceived as more important than characteristics concerning competence or sociability (as these are distinct dimensions of social judgment; Leach et al., 2007; in contrast to warmth in which both traits concerning morality and sociability are included; Fiske et al., 2007). Moreover, when people form an impression of a person or a group, they are more interested in information concerning morality traits than traits concerning competence and sociability (Brambilla et al., 2011; Brambilla et al., 2011). Indeed, when people form a first impression within milliseconds, they are more efficient in making inferences about trustworthiness than in making inferences about competence or likeability (Willis & Todorov, 2006).

People seem to be aware that moral judgments are important. For instance, Ellemers et al. (2008) demonstrated that people are inclined to adapt their choice to increase outcomes for the self or for the group to what other group members see as moral than to what other group members see as competent. Moreover, people anticipate being respected by their group members when they adjust their behavior to what the group considers moral (Pagliaro et al., 2011). These findings suggest that morality is of great importance for impression formation and deliberate impression management. We argue that people might also be more inclined to adjust their less deliberate actions (i.e., their implicit behavior) to what is considered moral than to what is considered competent.

In the current study, we examine this prediction using an Implicit Association Test (IAT) that is framed as a test of either individual morality or competence. The

IAT (Greenwald et al., 1998) has been used to measure implicit attitudes towards particular social groups, for example people with dark/white skin or, as in the current study, Muslim/non-Muslim women (see Appendix A). Targets in an IAT consist of stimuli representing social groups that are associated with positive and negative attributes. When people associate stimuli that represent their own (in-) group with positivity and stimuli that represent another (out-)group with negativity, they should respond more quickly and easily to trials that are congruent with these implicit associations than to incongruent combinations (e.g., ingroup stimuli and negative attributes). The IAT assesses the degree to which this is the case, as an indicator of implicit bias.

Whether IAT performance is really implicit and thus uncontrollable is much debated, however. There is much research showing the malleability of implicit attitudes, for example by repeated exposure to admired and disliked individuals (Dasgupta et al., 2009), emotions (Dasgupta & Greenwald, 2001), and several self and social motives (for a review see Blair, 2002). Moreover, it has been shown that the IAT effect is enhanced under stereotype threat (Frantz et al., 2004), but can be diminished when participants have a strategy that helps them to reduce their bias (Fiedler & Bluemke, 2005). In the current research we take advantage of the malleability of the IAT effect: We emphasize the social implications of participants' task performance (i.e., concerning their morality or their competence) and expect that participants to whom the implications concerning morality are emphasized will reduce their negative bias towards Muslim women. More specifically, we hypothesize that these participants will try to inhibit their implicit associations between Muslims and negative attributes, resulting in increased reaction times on congruent trials and thus a smaller IAT effect (consistent with the research of Fiedler & Bluemke [2005]).

Moreover, we are interested in the cognitive processes underlying the motivation to be moral and thus the inhibition of a negative bias on the IAT. Are intentions to behave in line with moral values associated with control of undesirable behavior, or do they influence selective attention that facilitates correct behavior? The current research addresses these questions using event-related brain

potentials (ERPs) associated with perceptual processing and conflict- and error monitoring.

Perceptual Attention

ERPs that are associated with early perceptual processing and more specifically, with selective attention and social categorization are the N1, the P150, and the N2. These components are associated with attention in such a way that increased amplitudes reflect the extent to which attention is directed towards a particular stimulus (e.g., Ito & Urlans, 2003). Moreover, research has shown that this attention differs between different social stimuli. For instance, the N1 – a negative deflection occurring around 100 ms after a stimulus is presented – is often larger when viewing stimuli resembling outgroup compared to ingroup members (i.e., black vs. white faces; Ito & Urland, 2003; Kubota & Ito, 2007; however, see Ito & Urland, 2005 for the reversed pattern). The P150, a positive peak that occurs somewhat later (approximately 150-250 ms post-stimulus, therefore also referred to as the P200), is also larger in amplitude for outgroup than for ingroup members (Ito & Urland, 2003; Kubota & Ito, 2007). In contrast to the N1 and P150, the N2 – a negative deflection around 200 ms post-stimulus – is found to be greater for stimuli representing the ingroup compared to the outgroup (Dickter & Bartholow, 2007; Ito & Urland, 2003, 2005). Examination of these components can thus show whether the emphasis on morality attracts greater attention to the group membership of the faces presented in the IAT (which is of importance when the test is said to measure participants' moral values concerning egalitarianism, but not when the test is said to measure competence). Moreover, components related to selective attention and social categorization can also be associated with motivated perception (e.g., the P150/P200; Amodio, 2010). We propose that emphasizing morality increases the motivation to suppress bias towards the outgroup. Although this could lead to diminished social categorization, we hypothesize that social categorization is actually enhanced: People's focus on the different group members should be increased to be sure to respond in line with egalitarian values (i.e., to be able to control implicit bias, as is also seen in research by Amodio, 2010). In other words, we expect to find stronger group-related modulations of the N1, P150, and N2 in the morality condition than in the competence condition.

Conflict- and Response Monitoring

Because we expect that emphasizing morality motivates people to inhibit their bias, we are also interested in ERPs associated with control. More specifically, conflict- and response monitoring. To assess conflict monitoring, we measure the N450. This is a negative modulation of the ERP signal, typically occurring around 400 ms post-stimulus, when subjects perform incongruent trials. The N450 modulation has been proposed to reflect the occurrence of response conflict (e.g., Nigam et al., 1992; Rebai et al., 1997), and is also evident in incongruent IAT trials (Williams & Themanson, 2011). Because the importance of trial congruency in the IAT may be more evident in the moral than in the competence IAT, and because we expect that control is increased when morality is made salient, we predict that the N450 modulation is larger in the morality compared to the competence condition.

To examine error-monitoring, we assess the error-related negativity (ERN), (Gehring et al., 1993; Nieuwenhuis et al., 2001). The ERN is a negative peak occurring within 100 ms after an erroneous response. The amplitude of the ERN is sensitive to the significance of errors. Hajcak et al. (2005) showed, for example, that ERN amplitude was greater on error trials when fast and accurate responses were associated with a large reward, and when participants' performance was being evaluated by a research assistant. In the current study, we hypothesize that subjects will be more motivated to prevent errors in the morality condition than in the competence condition, because the former might be viewed as a sign of immoral behavior, which is seen as more diagnostic for people's impression formation than incompetent behavior (Skowronski & Carlston, 1987). We therefore predict that erroneous responses will be associated with larger ERN modulations in the morality than in the competence condition.

We conducted two studies to test these predictions. In Study 2.1, we examined our hypothesis that social bias in the IAT is reduced when the test is said to measure participants' morality as opposed to their competence. In Study 2.2, we examined the cognitive processes associated with this reduced bias, as manifested in the ERP components discussed above.

Study 2.1

Method

Participants.

Sixty-six non-Muslim students from Leiden University (24 males, M_{age} = 20.2 years, *SD* = 1.8) participated in this study for money or course credits.

Procedure.

After providing written informed consent, participants performed five blocks of the IAT (Greenwald et al., 1998). Stimuli representing the target concepts consisted of 10 pictures of women without a headscarf (i.e., ingroup pictures) and 10 pictures of women with a headscarf (i.e., outgroup pictures; for details concerning the pretest of these stimuli, see Appendix A). Stimuli that represented positive and negative attributes consisted of 5 pictures of positive scenes, and 5 pictures of negative scenes, selected from the International Affective Picture System (Lang et al., 2005). The stimuli were selected based on the scores for pleasure (i.e., negative pictures with scores < 4 and positive pictures with scores > 7).

In a block of congruent trials, ingroup pictures shared the same response key as positive pictures and outgroup pictures the same response key as negative pictures. In a block of incongruent trials, this was the case for ingroup and negative pictures, and outgroup and positive pictures. The order of the congruent and incongruent blocks was counterbalanced across participants. Training blocks (IAT steps 1, 2 and 4) consisted of 26 trials, test blocks (steps 3 and 5) of 156 trials each. Every trial started with a fixation point (with a duration that varied between 500-1500 ms), followed by stimulus presentation (680 ms), and a feedback screen (500 ms). This screen indicated whether participants' response was correct (i.e., green check mark), incorrect (i.e., red cross), or "too late". Participants could not correct their incorrect responses.

Morality vs. competence task instruction. Participants were randomly assigned to an instruction condition. In the morality condition, participants read that the test would indicate their *values* concerning equal treatment of different people. In the competence condition, participants read that the test would indicate how well they are *able* to process new information (for the complete instructions,

see Appendix A). All participants were instructed to respond as quickly and accurately as possible. The test implications were repeated before the start of each test block.

Checks. To check that the perceived validity of the IAT did not differ between the conditions, we asked participants after they finished the test to respond to the statement: "My test score can assess what kind of person I am". Furthermore, two items measured participants' task engagement: "I think it is important to perform well on this test" and "It does not matter to me what my test score is" [reverse coded], (r = .62, p < .001). Participants could respond to each statement on a 7-point scale ranging from "completely disagree" (1) to "completely agree" (7). The experiment took approximately one hour after which participants were debriefed and thanked.

The IAT effect.

The dependent measure was the IAT effect, indicated by the *D* score. Based on the scoring algorithm described by Greenwald et al. (2003), this was calculated as the difference in reaction times on incongruent and congruent trials divided by a pooled *SD* of all correct trials. We included all trials, replaced error latencies with a replacement value (M + 2 $SD_{correct}$) and replaced latencies exceeding the maximum response time with the maximum response time of 680 ms.

Results and Discussion

Checks.

As intended, participants in the morality and competence condition did not think differently about the perceived validity of the test; M(morality) = 3.12, SD = 1.65; M(competence) = 3.24, SD = 1.60; F(1,64) < 1. Neither did they differ in their self-reported task engagement: M(morality) = 4.14, SD = 1.00; M(competence) = 4.24, SD = 1.16; F(1,64) < 1.

IAT effect.

Participants showed the standard IAT effect: A negative implicit bias towards the outgroup (i.e., women with a headscarf); t(65) = 4.72, p < .001. However, this bias was stronger in the competence condition; t(32) = 5.40, p < .001, than in the morality condition; t(32) = 1.77, p = .09. More importantly, an ANOVA predicting the *D* score from instruction conditions and order of test blocks revealed that the

bias was reduced in the morality, compared to the competence condition; *M*(morality) = 0.13, *SD* = 0.43; *M*(competence) = 0.34, *SD* = 0.36; *F*(1,62) = 4.56, p = .04, η^2 = .07. The reduced IAT effect was caused by a smaller difference between response times on incongruent and congruent trials in the morality condition: Consistent with previous research (Fiedler & Bluemke, 2005), participants in the morality condition responded somewhat more slowly on congruent correct trials than participants in the competence condition; *F*(1,64) = 3.24, p = .08 (see Figure 2.1)[1]. The percentages of errors did not differ between conditions; *M*(morality) = 8.81, *SD* = 6.03; *M*(competence) = 7.73, *SD* = 4.98; *F*(1,64) < 1. These behavioral results confirmed our hypothesis that task performance is adjusted when morality is made salient. To test which cognitive processes were modulated to produce the corresponding reduction in IAT score, we conducted Study 2.2.

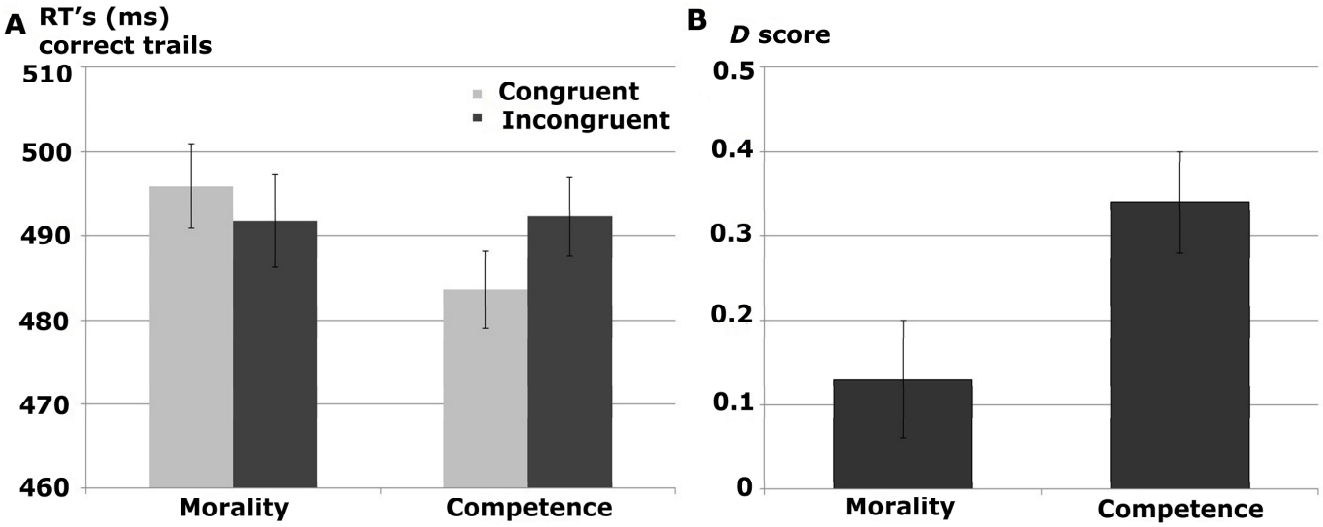

Figure 2.1. Reaction times (in milliseconds) on correct congruent and incongruent trials (A) and the IAT effect, in which error and missed trials are included after they are given a replacement value (*D* score; Figure B). Note that the reaction times on incongruent trials are quite fast relative to other IAT studies. This is caused by the limited the presentation time of the stimuli (i.e., participants had to respond within 680ms).

[1] We did not find decreased response times on incongruent trials (which could be expected based on conflict monitoring theory; e.g., Botvinick et al., 2001) because participants had a limited response time.

Study 2.2

Method

Participants.

Forty-four, healthy, right-handed, non-Muslim students from Leiden University (5 males, M_{age} = 20.4, SD = 4.3) provided written informed consent and participated in this study for money or course credits. One participant (morality condition) was excluded from the study due to an outlying IAT score; two participants (morality condition) had to be excluded from EEG analyses because of technical problems. Two more participants (one in each condition) were excluded from statistical analyses of the ERN because they did not make enough errors to reliably quantify this component (< 15).

Procedure.

Participants performed the IAT as described in Study 2.1, with the following modifications: We inserted a blank screen after the stimulus presentation to ensure that the ERN modulation occurred before the feedback. Each trial thus consisted of a fixation point (500 ms), a stimulus (680 ms), a blank screen (500 ms), and a feedback screen (750 ms). We also increased the number of congruent and incongruent trials from 156 to 300 to enhance the possibility that participants made enough errors to compute a reliable average ERN.

Participants' task engagement was measured with the items from Study 2.1 (r = .59, p < .001), and we checked whether participants in the morality condition were – as intended – more concerned about the social implications of their performance than participants in the competence condition (i.e., "I am concerned about the impression people might get of me, if they know how I performed on this test"). Moreover, we assessed the internal motivation to respond without prejudice (IMS) scale developed by Plant and Devine (1998; 5 items, α = .73; e.g., "I attempt to act in nonprejudiced ways toward women who wear a headscarf because it is personally important to me"; 7-point scale 1 "completely disagree" -7 "completely agree"). Previous research has shown that this internal motivation influences people's ability to regulate biased behavior by conflict-monitoring processes associated with the ERN (Amodio et al., 2008). Thus, to test our prediction that conflict- and error monitoring is enhanced in the morality

compared to the competence condition, we controlled for individual differences in IMS. The total experiment lasted 90 minutes, after which participants were debriefed and thanked.

EEG acquisition.

The EEG was recorded from 19 Ag/AgCl scalp electrodes, and from the left and right mastoids, using a 19-channel Biosemi active-electrode recording system (sampling rate 256 Hz). To assess horizontal and vertical eye movements, electrodes were placed on the outer canthi of the left and right eyes and approximately 1 cm above and below the right eye. EEG activity was recorded using ActiView software, offline data analyses were performed using Brain Vision Analyzer (BVA), and the experiment was controlled by E-prime (v 2.0). The EEG signal was referenced off-line to the average mastoid signal, corrected for ocular and eye-blink artifacts using the method of Gratton et al. (1983), and filtered (1-15 Hz). Single-trial stimulus-locked and response-locked epochs were extracted, ranging from -300 ms to 1000 ms after the event. These epochs were subjected to artifact rejection, then averaged and baseline-corrected by subtracting the average signal value between 200-0 ms pre-stimulus or between 300-50 ms prior to the response. Separate stimulus-locked ERP epochs were created for correct trials with outgroup and ingroup pictures, separately for the congruent and incongruent blocks. Separate response-locked ERP epochs were created for correct and incorrect responses. In an initial analysis, we found no effect of congruency on the ERN. Because participants made few errors on congruent trials, we pooled the congruent and incongruent trials to increase the number of trials averaged for each participant and thus the number of participants included in the ERN analysis.

ERP analyses.

Visual inspection of the data indicated that the N1, P150, N2, and N450 potentials were most evident at the midline electrode sites Fz, FCz, Cz, CPz, and Pz. These ERP components were quantified as the maximum peak amplitude within a time window (N1, 90-110 ms; P150, 100-250 ms; N2, 200-300 ms; N450 325-500 ms). To test the main effects of social categorization and conflict monitoring, we submitted the peak amplitude values to a 5 (electrode site) x 2

(picture type: ingroup/outgroup pictures) x 2 (congruency: congruent/incongruent trials) mixed-model ANOVA.

Visual inspection indicated that the error-related negativity (ERN) was largest at electrodes Fz, FCz, and Cz. To quantify the ERN, we determined the maximal (peak) amplitude of the signal between -50 and 150 ms around the response, separately for correct and incorrect trials. All peak amplitudes were submitted to a 3 (electrode site) x 2 (accuracy: correct/error) mixed-model ANOVA.

Because modulations of the task effects by the instruction manipulation were subtle, subsequent analyses focused on the electrode at which the interaction was most pronounced. The resulting peak-amplitude values were submitted to a mixed-model ANOVA with instruction condition as between-subjects variable and the relevant task factors as within-subject variables. Moreover, to control for individual differences in internal motivation to respond without prejudice, we included IMS score as a covariate in each analysis[2].

Results and Discussion

Checks.

As in Study 2.1, participants in the morality and competence condition did not differ in task engagement; M(morality) = 4.84, SD = 0.88; M(competence) = 4.63, SD = 0.94; $F(1,41) < 1$. Nor did they differ in their internal motivation to respond without prejudice; M(morality) = 4.89, SD = 0.82; M(competence) = 5.01, SD = 0.66; $F(1,41) < 1$. As expected, participants in the morality condition did report to be more concerned about the social implications of their performance than participants in the competence condition; M(morality) = 3.18, SD = 1.68; M(competence) = 1.91, SD = 1.02; $F(1,41) = 8.34$, $p = .006$, $\eta^2 = .17$.

Behavioral results.

Overall, participants showed the standard IAT effect (i.e., a negative implicit bias towards women with a headscarf); $t(42) = 5.04$, $p < .001$. Moreover, this bias was evident in both conditions; morality $t(20) = 2.52$, $p = .02$; competence $t(21) = 4.68$, $p < .001$. More importantly, an ANOVA with the D score based on the first 156 trials in each block as dependent variable, the instruction condition and the

[2] Inclusion of the IMS score only changed the results concerning the ERN, as is mentioned in the results section.

order of test blocks as independent variables, and IMS as covariate revealed a difference in the IAT effect between the instruction conditions: As in Study 2.1, the effect was smaller for participants in the morality condition than for participants in the competence condition; *M*(morality) = 0.13, *SD* = 0.40; *M*(competence) = 0.42, *SD* = 0.36; *F*(1,39) = 5.86, p = .02, η^2 =.13. As can be seen in Figure 2.2, this effect was caused by a smaller difference between response times on incongruent and congruent trials in the morality condition than in the competence condition. More specifically, (and similar to Study 2.1), participants in the morality condition responded somewhat more slowly on congruent trials than participants in the competence condition; *F*(1,41) = 3.06, p = .09. The percentages of errors did not differ between conditions; *M*(morality) = 12.36, *SD* = 7.13; *M*(competence) = 14.25, *SD* = 9.80; *F*(1,41) < 1. When we included all trials from each test block (a doubling of trials was needed for computing ERPs), the effect of condition was marginally significant; *M*(morality) = 0.15, *SD* = 0.27; *M*(competence) = 0.29, *SD* = 0.29; *F*(1,39) = 3.05, p = .09. This was caused by a training effect: Participants in both conditions responded faster and made fewer errors on the last 144 trials of each test block, resulting in a similar IAT performance. Although both analyses showed a main effect of the order of test blocks (respectively *F*[1,39] = 23.28, p < .001 and *F*[1,39] = 35.73, p < .001), this factor did not interact with instruction condition (*F*'s < 1).

ERP results.

Social categorization.

N1. We found the intended main effects of social categorization: The N1 was larger for outgroup pictures (*M* = -5.58 μV, *S.E.* = 0.32) than for ingroup pictures (*M* = -5.26 μV, *S.E.* = 0.30); *F*(1,38) = 6.86, p = .012, η^2 = .15. Analyses for the FCz electrode confirmed the predicted interaction between instruction condition and picture type; *F*(1,38) = 4.11, p = .050, η^2 = .10 (see Figure 2.3). The difference between the N1 elicited by outgroup and ingroup pictures was significant in the morality condition (*F*[1,38] = 4.69, p = .04, η^2 = .11), but not in the competence condition (*F*[1,38] < 1).

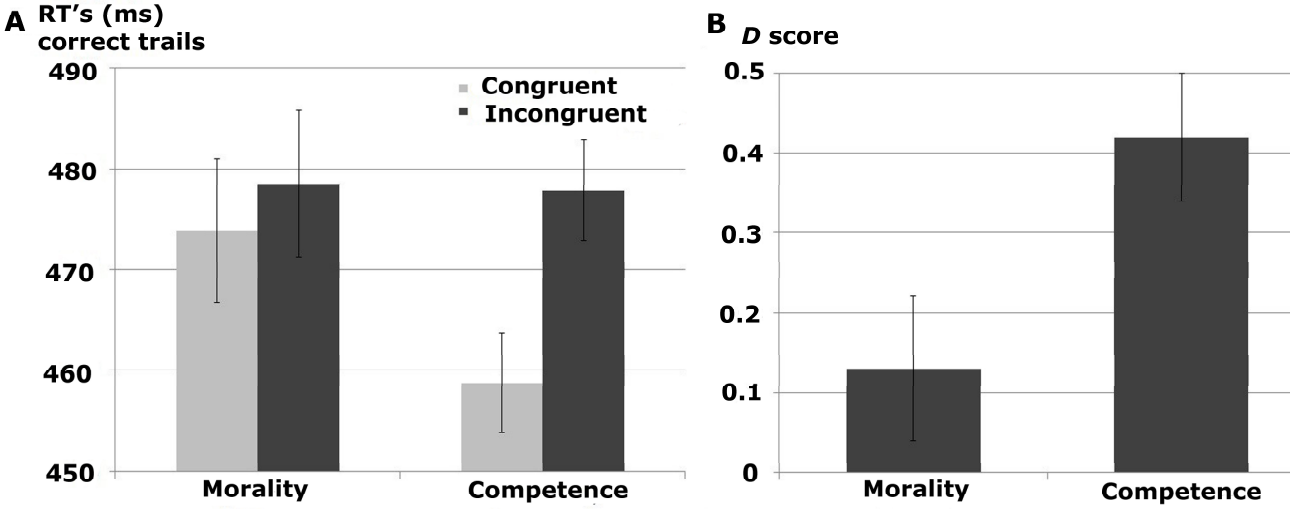

Figure 2.2. Reaction times (in milliseconds) on correct congruent and incongruent trials (A) and the IAT effect in which error and missed trials are included after they are given a replacement value (*D* score; Figure B). Note that the reaction times on incongruent trials are quite fast relative to other IAT studies. This is caused by the limited presentation time of the stimuli (i.e., participants had to respond within 680ms).

P150. As anticipated, the P150 was larger for outgroup pictures (M = 5.22 μV, *S.E.* = 0.52) than for ingroup pictures (M = 4.23 μV, *S.E.* = 0.52); $F(1,38) = 39.95$, $p < .001$, $\eta^2 = .51$. Analyses at Cz showed that, as predicted, there was an interaction effect between instruction condition and picture type; $F(1,38) = 5.12$, $p = .029$, $\eta^2 = .12$ (see Figure 2.3). The difference in P150 amplitude between outgroup and ingroup pictures was more pronounced in the morality condition ($F[1,38] = 33.75$; $p < .001$, $\eta^2 = .47$), than in the competence condition ($F[1,38] = 8.51$, $p = .006$, $\eta^2 = .18$).

N2. The N2 was, as intended, larger for ingroup pictures (M = -5.52 μV, *S.E.* = 0.50) than for outgroup pictures (M = -4.99 μV, *S.E.* = 0.47); $F(1,38) = 6.93$, $p = .012$, $\eta^2 = .15$. However, there was no interaction between picture type and instruction condition; $F(1,38) = 1.08$, $p = .31$.

Conflict- and error monitoring.

N450. Overall, the N450 was larger for incongruent trials (M = -2.22 μV, *S.E.* = 0.39) than for congruent trials (M = -1.45 μV, *S.E.* = 0.34); $F(1,38) = 12.51$, $p =$

.001, η^2 = 0.24. Analyses for the CPz electrode confirmed our prediction: Instruction condition interacted with congruency; $F(1,38) = 4.79$, $p = .035$, $\eta^2 = 0.11$ (see Figure 2.4). The difference in N450 amplitude between incongruent and congruent trials was significant in the morality condition ($F[1,38] = 16.12$, $p < .001$, $\eta^2 = .30$), but not in the competence condition ($F[1,38] = 1.20$, $p = .28$).

ERN. As anticipated, the ERN was larger for error trials (M = -6.83 μV, *S.E.* = 0.77) than for correct trials (M = 1.00 μV, *S.E.* = 0.53); $F(1,36) = 129.08$, $p < .001$, $\eta^2 = 0.78$. Moreover, accuracy interacted with IMS score; $F(1,36) = 4.03$, $p = .05$, $\eta^2 = .10$: A higher internal motivation to respond without prejudice was associated with larger ERN modulations (B = -1.46, $p = .09$; see also Amodio et al., 2008). However, more relevant to our current predictions, analyses at Cz showed a marginally significant interaction between accuracy and instruction condition; $F(1,36) = 3.49$, $p = .070$, $\eta^2 = .09$ (see Figure 2.5)[3]. The difference in ERN amplitude between error and correct trials was somewhat larger in the morality condition (M = -11.22 μV, *S.E.* = 1.17; $F[1,36] = 94.17$, $p < .001$, $\eta^2 = .72$) than in the competence condition (M = -8.38 μV, *S.E.* = 1.08; $F[1,36] = 59.74$, $p < .001$, $\eta^2 = .62$).

The ERP results are consistent with our expectations that stressing moral implications of the IAT increases social categorization of stimuli and conflict monitoring during the test. More specifically, the emphasis on morality moderates the attention towards outgroup but not ingroup faces (as indexed by increased N1 and P150, but not N2 modulations), and increases the neural response to response conflict and errors in the IAT (as reflected in increased N450 and ERN modulations), suggesting that erroneous responses were perceived as more significant in the morality than in the competence condition.

---

[3] The analysis without IMS as a covariate revealed the same pattern of moderation, but resulted in a non-significant interaction; $F(1,37) = 2.57$, $p = .12$. Moreover, as was put forward by an anonymous reviewer, the ERN results were sensitive to changes in the EEG processing settings. For example, shortening the baseline correction period (from 300-50 ms to 200-50 ms prior to the response) reduced the interaction effect between the ERN modulation and instruction; $F(1,36) = 2.72$, $p = .11$, $\eta^2 = 07$; whereas lowering the cutoff score for the high-pass filter (from 1 to 0.1 Hz) made this interaction significant; $F(1,36) = 4.97$, $p = .03$, $\eta^2 = .12$.

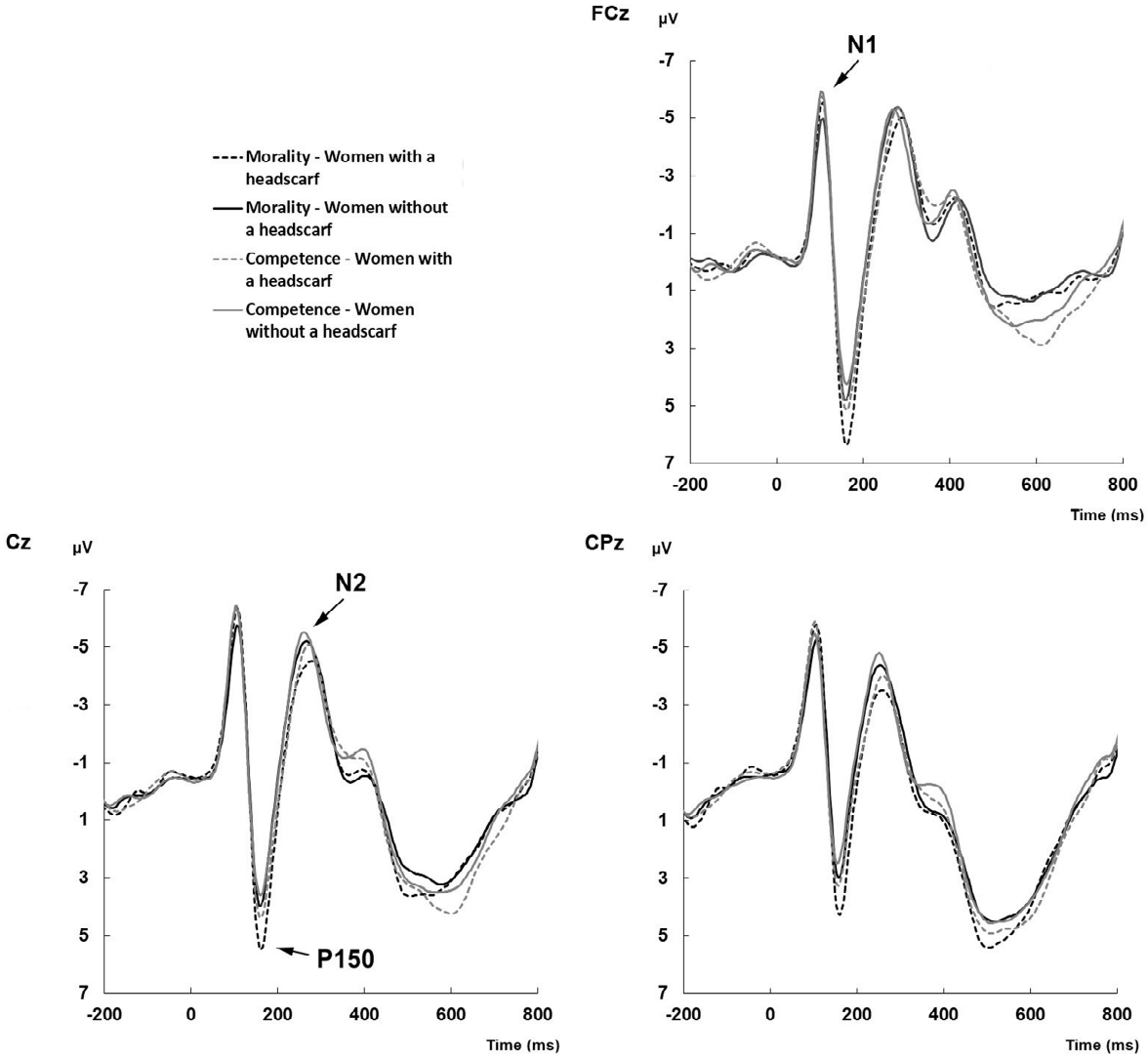

Figure 2.3. The N1, P150 and N2 modulations for pictures of women with and without a headscarf at three central electrodes. The interaction with instruction condition was significant at FCz for the N1, and at Cz for the P150. The interaction did not reach significance for the N2.

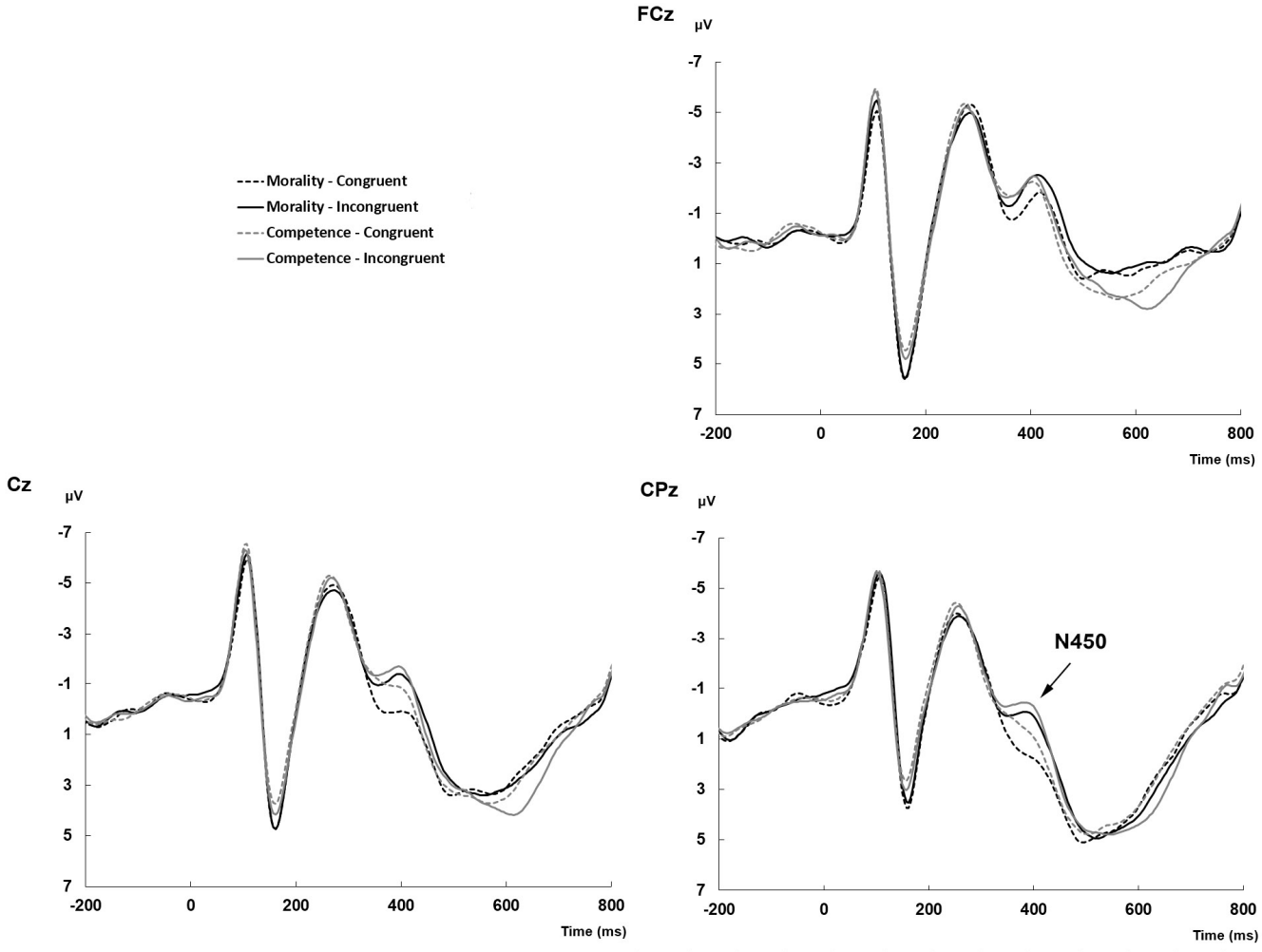

Figure 2.4. The N450 modulations for incongruent and congruent trials at three central electrodes. The interaction with instruction condition was significant at CPz.

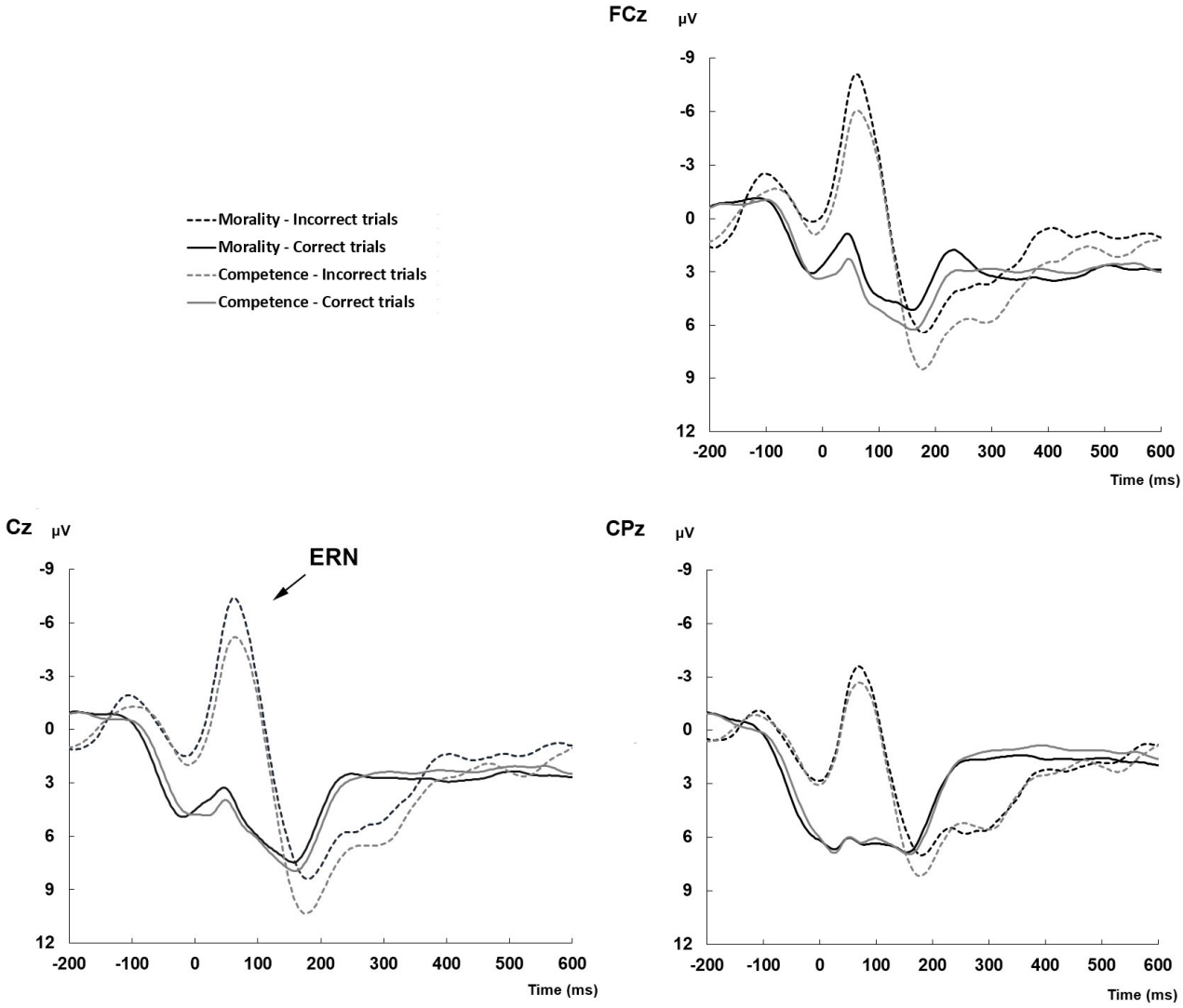

Figure 2.5. The ERN modulations for correct and incorrect trials at three central electrodes. The interaction with instruction condition was marginally significant at Cz.

General Discussion

Previous research has shown that morality is more important than competence for people's personal and social identity (e.g., Leach et al., 2007), and that morality guides explicit strategic behavior (Ellemers et al., 2008). The present studies extend prior research by showing that morality also impacts on non-explicit aspects of task behavior: People inhibited their negative bias towards Muslim women on an IAT when the test was said to be indicative of their morality (instead of their competence). Our findings thus reveal that participants are able to reduce their implicit bias when given the opportunity to reveal their moral side. This complements prior observations that implicit bias is exacerbated when participants are identified as potential racists (Frantz et al., 2004), and is consistent with research showing that moral appeals induce different physiological and behavioral responses, depending on whether these are framed as ideals or as obligations (Does et al. 2011; 2012).

Importantly, the current research provides insight into the neurobiological mechanisms underlying the differential performance on the moral and competence IAT. Previous research has shown that performance on tasks designed to measure implicit attitudes are associated with (increased) motivated perception (Amodio, 2010) and response monitoring (Amodio, et al., 2008). Additionally, this study reveals that these cognitive processes are activated or enhanced when people's morality is emphasized. More specifically, when morality is emphasized as opposed to competence, people engage in increased social categorization of outgroup faces, and in enhanced conflict- and response monitoring. Because these processes have previously been associated with motivational states (e.g., Amodio, 2010; Hajcak et al., 2005) and because morality has been shown to be more important than competence for impression formation and -management, we interpret these findings as indicating increased motivation of participants in the morality condition to control their bias on the IAT.

The findings concerning increased conflict- and error monitoring during a moral IAT also extend research showing that low levels of implicit bias (often revealed by people with high internal and low external motivation to avoid prejudice) are associated with successful response monitoring (Amodio et al., 2008;

Gonsalkorale et al., 2011). The current results additionally indicate that, regardless of individual differences in internal motivation to respond without prejudice, emphasizing moral values successfully reduces displays of implicit bias. Moreover, our results indicate that emphasizing morality affects not only corrective processes like error monitoring, but affects performance through processes involved in the attention to social stimuli before responses are given.

Although the current research broadens the knowledge of the importance of morality for people's self-identity, we also mentioned that morality is more important than competence for people's *social* identity, and their behavior in groups (Ellemers, et al., 2008; Leach et al., 2007). The question thus remains whether our findings would be affected by for example social evaluation. Further research could address this question by examining whether the emphasis on morality influences people's task performance in the presence of other people and whether this differs between evaluations of ingroup compared to outgroup members.

Conclusion

Our findings extend previous research that demonstrates the importance of morality over competence for people's self-view. In particular, our findings show that people control their implicit responses during a moral task, and reveal how they do that: Emphasizing morality facilitates people's task performance by increasing perceptual attention and conflict- and error monitoring.

Acknowledgements

We thank Ilona Domen, Suzanne Cederhout, Reinier Lagerwerf, Piarella Rodriguez, Lenny van den Beukel, Jelle van Hasselt, and Bart van Wingerde for their help with the data collection, and David Amodio, Guido Band, Stephen Brown, Eveline Crone and Henk van Steenbergen for their advice.

Chapter 3

Emphasizing moral task implications influences visual attention: An fMRI study

Collaborators on the research described in this chapter are:
Naomi Ellemers, Anna van Duijvenvoorde, Belle Derks, Eveline Crone, and Serge Rombouts. Their contribution can be specified as follows: Design of the study: FvN, NE, EC. Performing the experiment: FvN. Data analyses: FvN, AvD. Additional data processing: EC, SR. Writing of the paper: FvN, NE, AvD, BD.
Manuscript in preparation.

Morality, having the knowledge of and behaving according to what is right and wrong, is seen as key to human social life: It is one of the hallmarks of society since it is the basis for people's individual choices, their social interactions and group functioning. To gain understanding of how (im)moral behavior is initiated, much research has focused on the development of people's individual level of rational decision making: Knowing what is and what is not moral and considering what would be the best thing to do in particular situations. Using neuroscientific research methods, researchers have been able to reveal the neural networks involved in such moral cognition. Specifically, participants in those studies are often asked to take the observers perspective and judge the (im)moral content of phrases or pictures (e.g., Cope et al., 2010), to decide on different moral dilemmas (e.g., Christensen & Gomila, 2012), or to imagine behaving in line with or opposed to moral norms (e.g., Decety & Porges, 2011). However, as Casebeer (2003) noted, thinking about (doing) moral things is different from actually doing moral things –and imagined compared to real moral decision making is even associated with different neural networks (FeldmanHall et al., 2012). In the current research, we therefore aimed to extend previous research on moral cognition by examining how people's motivation to behave morally affects their actual performance on a task said to be indicative of their moral values. Moreover, we investigated how such moral motivation affects the cognitive processes involved in this task performance.

In neuroscientific research on moral psychology the social significance of morality is often underemphasized or even excluded (Casebeer, 2003). Nevertheless, moral choices and behaviors are inherently social: They often imply taking care of others or treating others well. In fact, some analyses consider morality and sociability as representing one evaluative domain, although they encompass different characteristics and behaviors (Leach, Ellemers, & Barreto, 2007). Indeed, judging other people's moral integrity and trustworthiness is important in social interactions (e.g., Brambilla, Rusconi, Sacchi, & Cherubini, 2011; Delgado, Frank, Phelps, 2005). Moreover, being perceived as a moral person is important for one's social identity: People experience pride in being a member of a group with high morality (Leach et al., 2007), and they indicate a willingness to adhere to moral group norms (Ellemers, Pagliaro, & Barreto, 2008) because they

expect to receive respect from their fellow group members in this way (Pagliaro, Ellemers, & Barreto, 2011). Being moral thus encompasses more than intrapsychological processes associated with cold moral reasoning. Even when we know the moral thing to do and certain brain regions may be associated with coming to that decision, it is not self-evident that such mechanisms are also associated with actual behavior. It is therefore important, in addition to the investigation of moral cognition, to increase our understanding of the neural processes involved in the motivation to display moral behavior.

Prejudice Control as an Indicator of Moral Behavior

As mentioned above, being moral often has social implications: Defining what is right or wrong may depend on what others value as the moral thing to do and on how others are affected by our actions. In the current research, we examine moral behavior in the context of intergroup relations and prejudice: Fairness towards and the equal treatment of different groups in society are seen as core moral values. There thus tends to be a general motivation to be or to appear unprejudiced. Because of those moral and social implications, prejudice is often measured on an implicit level, for instance with an implicit association test (IAT). The IAT (Greenwald, McGhee, & Schwartz, 1998) was first designed to assess people's positive versus negative associations with particular social groups: Their implicit social bias. Stimuli in this reaction time test consist of target concepts –representing members of social groups, such as faces of Black and White men, or Muslim and non-Muslim women– and positive and negative attributes. On prejudice-congruent trials, participants are asked to categorize the stimuli representing their own (in-)group using the same response key as positive attributes, and stimuli representing another (out-)group and negative attributes with another key. On prejudice-incongruent trials they are asked to categorize stimuli representing their ingroup and negative attributes with the same key, as well as stimuli representing the outgroup and positive stimuli. To the extent that people are more inclined to associate their ingroup with positivity and the outgroup with negativity, they should respond more quickly and easily to the congruent as compared to the incongruent trials. The IAT assesses this difference in response latencies on incongruent compared to congruent trials, as an indicator of implicit bias.

Recent research has revealed that people are able to influence their performance on an IAT if they are motivated to do so. For instance, Fiedler and Bluemke (2005) have shown that participants can reduce their negative bias when they are aware of how the IAT bias is computed and when they are encouraged to find out effective strategies to adjust their performance. Moreover, Van Nunspeet, Ellemers, Derks, and Nieuwenhuis (2014) showed that people's motivation to control prejudice was higher when the moral implications of the IAT were emphasized, resulting in a weaker bias against Muslim women.

In the current research, we will adopt the same paradigm as used by Van Nunspeet et al. (2014) to create circumstances that amplify the motivation to behave morally (i.e., to control expressions of implicit bias). In addition, we use functional magnetic resonance imaging to examine the neural processes underlying such moral behavior.

Neural Correlates of Social Bias Control

In their study, Van Nunspeet et al. (2014) measured brain activation with an electroencephalogram (EEG) to examine the cognitive processes underlying moral IAT performance. Their results revealed that participants who had been reminded of the moral implications of the IAT (as compared to a control condition) showed increased perceptual attention to the different types of targets in the IAT (both in terms of group membership and individuating facial features, as indicated by increased N1 and P150 modulations in response to viewing the pictures of outgroup and ingroup targets; Van Nunspeet et al., 2014). The present study aims to further examine these processes using fMRI, by examining how the motivation to perform in line with one's moral values affects patterns of brain activation associated with performance on an IAT.

In fMRI research, face perception is often located in the inferior part of the occipital lobe. More specifically, within the inferior occipital gyrus (also called the occipital face area, OFA; for a review see Pitcher, Walsh, & Duchaine, 2011) and the fusiform gyrus (FG; but see Haxby, Hoffamn, & Gobbini, 2000; and Ishai, 2008, for more complete overviews of the cortical network involved in face processing). Activation in the OFA is associated with facial recognition (i.e., the establishment that a face is a face), which occurs at an early stage of visual

perception (Pitcher et al., 2011). In contrast, activation in the FG is associated with the subsequent and deeper processing of higher-level facial features. For example, activation in the FG is greater when people view ingroup compared to outgroup members (e.g., Kubota, Banaji, & Phelps, 2012; Van Bavel, Packer, & Cunningham, 2011). Since previous research revealed that when the moral implications of the IAT are emphasized, perceptual attention towards and social categorization of ingroup and outgroup faces is increased (Van Nunspeet et al., 2014), we hypothesized that participants who performed the moral (as compared to the control) IAT in the current research would show increased activation in the FG when viewing ingroup as compared to outgroup targets. Moreover, since the process of social categorization was found in early event-related brain potentials (i.e., around 100 and 150ms after stimulus-onset), we also wanted to examine whether we could find any evidence for increased social categorization of ingroup and outgroup targets in the OFA given its association with early facial processing.

Another finding in the EEG study was that the inhibition of social bias on the IAT was associated with increased modulations of the error-related negativity (ERN; Van Nunspeet et al., 2014). The ERN is associated with response-monitoring (e.g., Gehring, Goss, Coles, Meyer, & Donchin, 1993; Nieuwenhuis, Blom, Band, & Kok, 2001) and the significance of making errors (Hajcak, Moser, Yeung, & Simons, 2005). Specifically, results of the research of Van Nunspeet et al. (2014) showed that conflict- and error-monitoring was enhanced for participants to whom the moral implications of the IAT were emphasized, indicating the increased significance of making errors on a task indicative of their moral values.

The conflict- and response monitoring processes found by Van Nunspeet et al. (2014) are in line with patterns of brain activation found in fMRI studies on social bias: In studies using the IAT, brain activation associated with performance on incongruent IAT trials is contrasted to brain activation associated with performance on congruent IAT trials. Results reveal that performance on the incongruent compared to congruent trials is associated with increased activation in the dorsolateral prefrontal cortex (dlPFC) and anterior cingulate cortex (ACC; e.g., Chee, Sriram, Soon, & Lee, 2000; Stanley, Phelps, & Banaji, 2008). These brain regions are known to be involved in conflict monitoring and control of behavior

(Botvinick, Braver, Barch, Carter, & Cohen, 2001; MacDonald, Cohen, Stenger, & Carter, 2000) and specifically, in the area of prejudice, considered as regulating (implicit) bias (Stanley et al., 2008). In the current research, we aim to extend the findings of Van Nunspeet et al. (2014), namely that participants tend to be highly vigilant while performing an IAT framed as a measure of their morality –as compared to a control condition in which the IAT is framed as a measure of their competence. Accordingly, we examined whether the motivation to perform in line with moral values affected brain activation in regions associated with cognitive control when participants perform incongruent versus congruent IAT trials.

Triangulation

Whereas cognitive processes associated with people's concerns to behave in line with their moral (e.g., egalitarian) values have been revealed in previous EEG research, our current goal is to expand these insights using fMRI. Both methodologies have their advantages: EEG has a high temporal resolution, making it is possible to examine the onset and time course of different cognitive processes, including very early and immediate responses. In addition, fMRI has a high spatial resolution which gives us the opportunity to locate the brain regions involved in moral task performance. Thus, whereas previous research has revealed that perceptual attention to different types of faces is increased (as seen in the N1 and P150 potentials, measured at the frontocentral sites of the scalp; Van Nunspeet et al., 2014), the current research will examine whether this is also evident in patterns of brain activation in the visual cortex. Moreover, we can investigate whether the enhanced error-detection and conflict-monitoring processes found in EEG research are also evident in brain areas associated with cognitive control. By using such a triangular approach (i.e., combining insights from behavioral, EEG and fMRI research) we will get a better understanding of people's motivation to *be* moral, in addition to our knowledge of brain processes and networks involved and necessary for moral cognition.

Study 3

Method

Participants

Twenty-six, non-Muslim, right-handed students from Leiden University participated in the study. None of the participants reported a history of psychiatric or neurological disorders, or current use of any medications. One participant was excluded from the data because of a software failure during the scanning session; two other participants were excluded from the fMRI data analyses because of technical problems. The remaining twenty-three participants (8 males, M_{age} = 21.0 years, *SD* = 4.9) were randomly divided across the two conditions of the between-participants design (i.e., the morality [*N* = 11] or control [*N* = 12] task domain). All procedures were approved by the medical ethical committee of the Leiden University Medical Center (LUMC) and all participants gave informed consent for the study.

Morality Framing of the IAT

While in the scanner, before the start of the IAT, half of the participants read that the computer task they were going to perform could indicate their endorsement of *moral values* concerning egalitarianism and discrimination (the morality condition). The other half of the participants was informed that the test could indicate their *ability* to process new information and to learn new tasks (the control condition). All participants were instructed to respond as quickly and accurately as possible. They were reminded about the test implications before the start of each test run (i.e., runs 3 and 5; see also Van Nunspeet et al., 2014).

Instruments and Procedure

Participants performed the five steps (blocks) of the IAT as designed by Greenwald et al., (1998). We used an event-related block design: Each scanning run consisted of one IAT block, but each stimulus was preceded by a fixation cross creating a jittered interstimulus interval (min. = 1100 ms, max. = 6600 ms) in order to model the hemodynamic response function for each stimulus type. After the fixation cross, the stimulus was presented (with a maximum duration of 680 ms in which participants were asked to respond [e.g., Beer et al., 2008]; more details about the stimuli below) which was followed by a feedback screen (1650 ms minus

the reaction time on the stimulus; see Figure 3.1). The feedback screen consisted of a green checkmark (i.e., correct response), a red cross (i.e., incorrect response), or the words "too late" when the participant did not press a key on time. Only trials on which participants responded correctly and on time were analyzed.

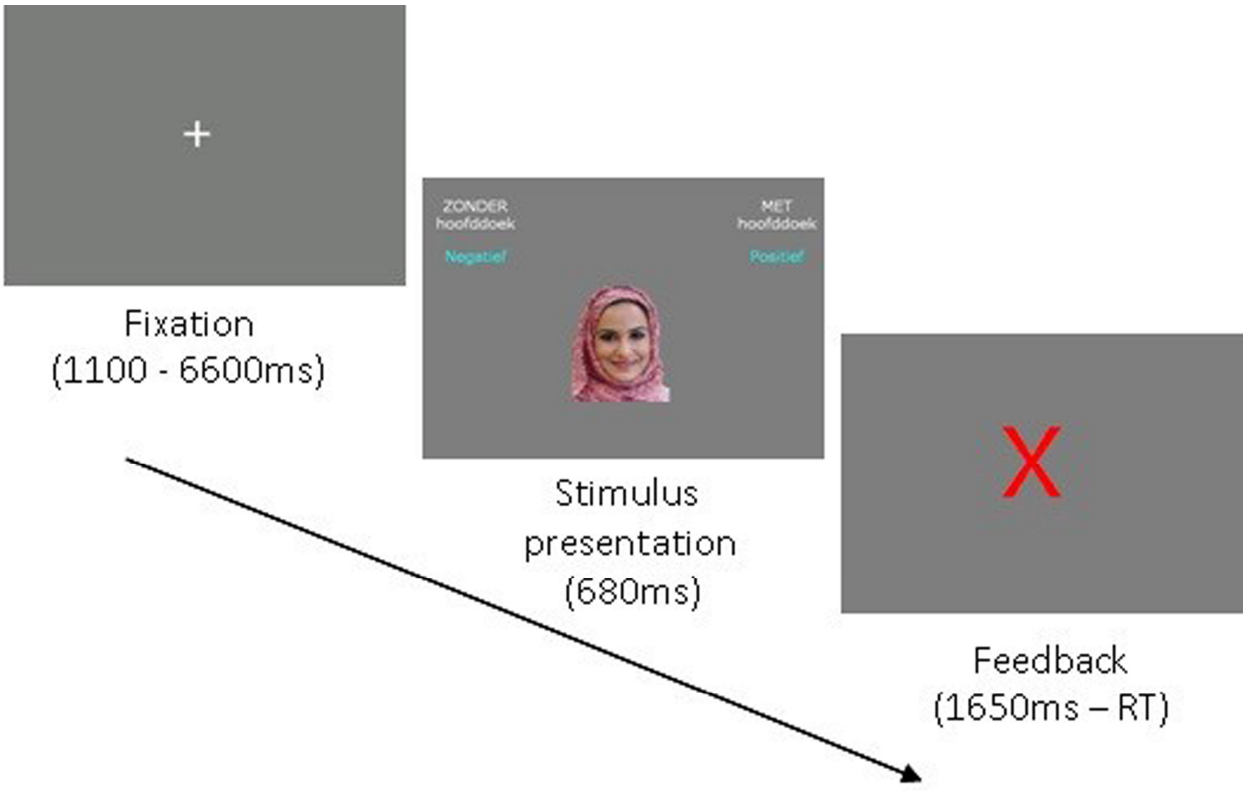

Figure 3.1. Example of an IAT trial. RT = reaction time.

In run 1, stimuli consisted of 10 faces of Muslim women (wearing a headscarf; outgroup pictures) and 10 faces of non-Muslim women (not wearing a headscarf; ingroup pictures) which participants were asked to categorize using a right (index finger) or left (index finger) response key. In run two, stimuli consisted of 5 pictures of positive scenes, and 5 pictures of negative scenes (International Affective Picture System; Lang et al., 2005). In run three, both picture types were presented and participants responded either with one key to outgroup pictures and negative scenes and with the other key to ingroup pictures and positive scenes (i.e., congruent trials). Or they responded with one key to outgroup pictures and positive scenes and with the other key to ingroup pictures and negative scenes (i.e.,

incongruent trials). Run four was similar to run one except for the fact that the response keys for the ingroup and outgroup pictures were switched. Finally, run five was similar to run three: Both ingroup/outgroup pictures and pictures of positive/negative scenes were presented. However, when congruent trials (i.e., 'ingroup + positivity' and 'outgroup + negativity) were presented in run3, then run 5 consisted of incongruent trials (i.e., 'outgroup + positivity' and 'ingroup + negativity'), and vice versa. The order of the runs was thus counterbalanced between participants. Training runs 1, 2, and 4 consisted of 20 trials each and lasted approximately two minutes. Testing runs 3 and 5 consisted of 120 trials each and lasted approximately six minutes. All IAT instructions were presented on the screen in the scanner bore before the start of each run. Since the experiment was part of a larger study, participants spent approximately 2 hours in the laboratory, and received 20 euros as a compensation for their participation.

fMRI Data Acquisition and Analysis

Scanning was performed at the Leiden University Medical Centre (LUMC) with a standard whole-head coil on a 3.0 Tesla Philips Achieva scanner. Using E-prime 1.0 software, the IAT was projected onto a screen at the back of the scanner bore, which participants could view via a window attached to the top of the head coil. Participants could respond by pressing keys on boxes attached to their legs. The IAT consisted of five event-related runs, of which we only analyzed test runs 3 and 5 (consisting of congruent and incongruent trials). Functional data were obtained using T2*-weighted echo-planar imaging ([EPI], repetition time (TR) = 2200 ms, echo time (TE) = 30 ms, slice matrix = 80 x 80, slice thickness = 2.75 mm, slice gap = 0.28 mm, field of view [FOV] = 220 mm). A high-resolution 3D T1-weighted anatomical image (TR = 9.751 ms, TE = 4.59 ms, flip angle = 8°, 140 slices, 0.875 mm x 0.875 mm x 1.2 mm, and FOV = 224.000 x 168.000 x 177.333) was collected at the end of the scanning session.

Data were preprocessed and analyzed using SPM8 software (Welcome Department of Cognitive Neurology, London) implemented in MATLAB (Mathworks, Sherborn, MA). The functional time series were realigned to compensate for small head movements. Translational movement parameters never exceeded 1 voxel (< 3 mm) in any direction for any subject or scan. Functional

volumes were spatially normalized to EPI templates. The normalization algorithm used a 12 parameter affine transformation together with a nonlinear transformation involving cosine basis functions and resampled the volumes to 3 mm cubic voxels. Functional volumes were spatially smoothed using an 8 mm full-width half-maximum Gaussian kernel. Templates were based on the MNI305 stereotaxic space (Cocosco, Kollokian, Kwan, Pike, & Evans, 1997), and the Montreal Neurological Institute (MNI) atlas was used to refer to the coordinates.

To analyze the data, a canonical hemodynamic response function was convolved at the onset of the stimulus and modeled as a zero-duration event. We distinguished between eight different types of stimuli: The IAT consisted of pictures of ingroup targets, outgroup targets, positive scenes, and negative scenes. Moreover, these stimuli were presented in a congruent as well as an incongruent run. Invalid trials were included in the model as a regressor of no interest. Two main contrast analyses were distinguished: To examine brain activation associated with visual perception of ingroup and outgroup targets, we investigated a contrast of viewing faces of non-Muslim women versus Muslim women, collapsed over the two IAT test runs (i.e., congruent/incongruent ingroup targets and congruent/incongruent outgroup targets, measured *within* participants). Moreover, to test whether this activation interacted with the emphasis on the moral implications of the task (measured *between* participants), we conducted a 2 (ingroup/outgroup targets) X 2 (morality/control) full factorial ANOVA.

To examine activity associated with the IAT effect (measured *within* participants), brain activation for the incongruent IAT run (for all ingroup / outgroup / positive / negative pictures) was compared to brain activation during the congruent IAT run (also for all ingroup / outgroup / positive / negative pictures). Moreover, to test whether the activation associated with the IAT effect interacted with the emphasis on the moral compared to the competence implications of the task (measured *between* participants), we conducted a 2 (incongruent/congruent) X 2 (morality / control) full factorial ANOVA.

The analyses were carried out using the general linear model in SPM8. For each individual, contrast parameter images were computed and the resulting contrast images were submitted to second-level group analyses. Only effects of at

least 10 continuous voxels that exceeded a False Discovery Rate (FDR) corrected threshold of $p < .05$ are reported.

Moreover, since we were interested in the –perhaps quite subtle– difference between the emphasis on the moral compared to the competence implications of the task, we also extracted parameter estimates from the regions of interest (ROI) that were identified in the whole brain analyses to explore the pattern of the activation across our conditions. These regions were extracted using the Marsbar toolbox (Brett, Anton, Valabregue, & Poline, 2002) for SPM8.

Results

Behavioral Results

The IAT effect is indicated by the *D* score, and measured as the difference in reaction times on incongruent and congruent trials divided by a pooled *SD* of all correct trials (Greenwald, Nosek, & Banaji, 2003). We included all trials, replaced error latencies with a replacement value ($M + 2\ SD_{correct}$; Greenwald et al., 2003) and replaced latencies exceeding the maximum response time with the maximum response time of 680 ms. The resulting positive *D* scores are an indication of people's evaluative bias against the outgroup (i.e., Muslim women).

To test whether participants showed an IAT effect overall, we conducted a one-sample t-test with *D* score as the dependent variable and a comparison test score of zero. As expected, results revealed the standard IAT effect; $M = 0.18$, $SD = 0.33$, $t(24) = 2.66$, $p = .01$, indicating that participants showed bias against Muslim women. Subsequently, we tested whether the task domain manipulation influenced the IAT effect. Specifically, whether emphasizing the moral implications of the IAT caused participants to show a smaller bias against Muslims. However, contrary to our hypothesis, an ANOVA with *D* score as dependent variable and task domain and the order of IAT test blocks as independent factors showed no main effect of task domain, nor an interaction effect between task domain and order; *F*'s < 1. There was only a main effect of order, $F(1,21) = 9.52$, $p = .006$, $\eta_p^2 = .31$, indicating that participants who performed the congruent block first showed a smaller bias against Muslim women ($M = 0.01$, $SD = 0.22$) than participants who performed the incongruent block first ($M = 0.36$, $SD = 0.34$). Perhaps, this effect is due to starting the task with the relatively difficult trials which could increase

response latencies and thus the difference between responses on incongruent and congruent trials.

Even though we observed no differences at the overt behavioral response level, it is still of interest to see whether different brain areas are involved in displaying these responses dependent on experimental conditions.

Imaging Results

Face perception.

To examine the neural activation associated with viewing faces of outgroup members (Muslim women) and ingroup members (non-Muslim women), we first conducted a 2 (Target identity: ingroup/outgroup faces) x 2 (Task Domain: morality/control) full factorial ANOVA at the whole brain level. Results revealed no main effects, nor an interaction. One-sample t-tests –averaged across the task domain conditions– showed no significant patterns of activation for the outgroup > ingroup targets contrast. However, as expected, the ingroup > outgroup targets contrast showed a significant difference in activation in the bilateral fusiform gyrus (see Table 3.1 and Figure 3.2A), indicating that –in line with previous research (e.g., Kubota et al., 2012; Van Bavel et al., 2011)– activation was greater when participants viewed faces of ingroup members (non-Muslim women) as compared to faces of outgroup members (Muslim women).

In addition to the whole-brain analyses, we extracted parameter estimates from the regions of interest (ROIs) that were identified in the whole brain analyses to further examine the patterns of activity between participants in the morality and the control condition. Specifically, we localized ROIs in two area's in the visual cortex known to be associated with processing faces: The fusiform gyrus (FG, Brodmann area 37) and the occipital face area (OFC, Brodmann area 19). Both ROIs were based on the contrast of 'All faces' (i.e., congruent/incongruent ingroup targets and congruent/incongruent outgroup targets) > 'fixation' (FDR corrected $p < .05$, 10 continuous voxels). Within this contrast, we located the FG and OFA bilaterally and the peaks of the activation (MNI coordinates FG: +39, -49, -26 and -36, -43, -26; MNI coordinates OFA: -36, -67, -20 and +42, -64, -23) defined the centers of four 10 mm diameter sphere-shaped ROIs (Figure 3.2; see Ratner, Kaul, & Van Bavel, 2013, for a similar approach). Parameter estimates from these ROIs

were included as the dependent variable in a 2 (Hemisphere: left/right) x 2 (Target identity: ingroup/outgroup faces) x 2 (Congruency: congruent/incongruent) repeated measures ANOVA with Task Domain (morality/control) and order of the IAT test blocks (congruent/incongruent first) as independent factors. Relevant to our interest in face perception, we did not find a main effect of, nor any interaction effects with task domain in the FG. Consistent with the whole-brain analysis, only the effect of target identity was significant indicating that activation in the FG was greater for viewing ingroup compared to outgroup faces, $F(1,19) = 7.49$, $p = .01$, $\eta_p^2 = .28$.

Results concerning face perception in the OFA also showed the main effect of target identity, $F(1,19) = 7.41$, $p = .01$, $\eta_p^2 = .28$, indicating that activation was greater for viewing ingroup as compared to outgroup faces. There was also an interaction effect between congruency and order, $F(1,19) = 4.29$, $p = .05$, $\eta_p^2 = .18$: For congruent trials, activation in the OFA was greater when the congruent (rather than the incongruent) run was presented first, $F(1,19) = 10.41$, $p = .004$, $\eta_p^2 = .35$. The other simple main effects were not significant, F's ≤ 2.39, p's $\geq .14$. Additionally and more interestingly, we observed a marginally significant interaction effect between target identity, congruency, and task domain, $F(1,19) = 3.56$, $p = .07$, $\eta_p^2 = .16$. To interpret this complex interaction, we conducted separate analyses for the control and morality conditions separately. Results revealed that there were no main effects of target identity or congruency, nor an interaction effect in the control condition; F's ≤ 2.39, p's $\geq .15$. However, in the morality condition, there was a significant main effect of target identity; $F(1,11) = 13.90$, $p = .003$, $\eta_p^2 = .56$, indicating greater activation for ingroup compared to outgroup targets. There was also a marginally significant main effect of congruency; $F(1,11) = 4.22$, $p = .07$, $\eta_p^2 = .28$, showing that activation in the OFA was greater on congruent compared to incongruent trials (see Figure 3.3). This findings is consistent with previous research, showing that participants adjusted their behavioral responses on prejudice-congruent trials when the moral implications of the IAT were emphasized (Van Nunspeet et al., 2014). There was no interaction effect; F's ≤ 1.62, p's $\geq .23$.

IAT effect. To examine the neural correlates of the IAT effect (i.e., bias against Muslim women), we examined neural activation associated with participants' performance on congruent versus incongruent trials. We first conducted a 2 (Congruency: incongruent/congruent) x 2 (Task Domain: morality/control) full factorial ANOVA at the whole brain level. Results revealed no significant main effects nor an interaction effect. Additionally, we conducted one-sample t-tests, averaged across the task domain conditions. Whole-brain contrasts (congruent > incongruent and incongruent > congruent) revealed no significant differences.

In addition to the whole-brain analyses, we again identified ROIs to further examine the patterns of activity between participants in the morality and the control conditions. Specifically, we localized ROIs based on activation in the contrast 'All stimuli' (i.e., congruent/incongruent ingroup targets; congruent/incongruent outgroup targets; congruent/incongruent positive scenes; and congruent/incongruent positive scenes) > 'fixation' (FDR corrected $p < .05$, 10 continuous voxels). Results of this contrast did not reveal activation in the ACC. However, we were able to detect activation within the right DLPFC and the peak of this activation (MNI coordinates: +45, +32, +28) defined the center of a 10 mm diameter sphere-shaped ROI. Parameter estimates from this ROI were included as the dependent variable in a repeated measures ANOVAs with congruency (i.e., congruent/incongruent trials, averaged over picture type) as within-participants factor and Task Domain (morality/control) and order of the IAT test blocks (congruent/incongruent first) as independent factors. Results of this analysis showed however no effects of congruency, task domain, or an interaction effect, F's ≤ 2.87, p's $\geq .11$.

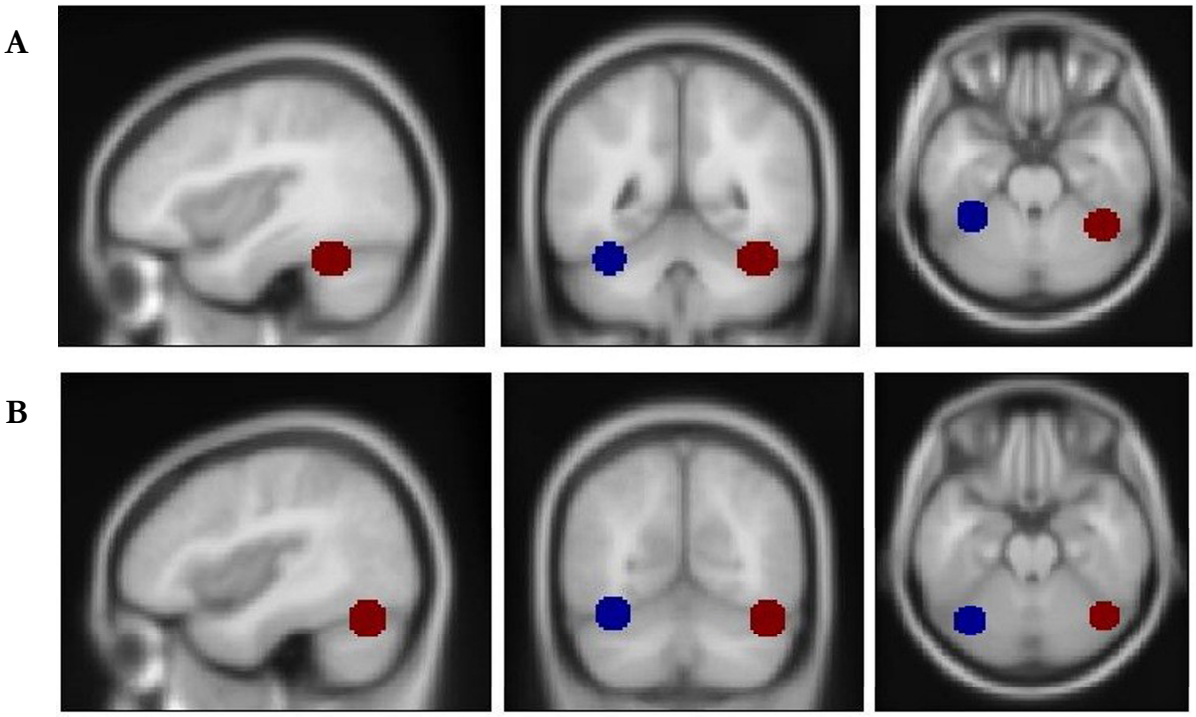

Figure 3.2. Activation was found in the bilateral fusiform gyrus (FG; Brodmann area 37) and occipital face area (OFA; Brodmann area 19) in the *faces* > *fixation* contrast (FDR corrected $p < .01$, 20 continuous voxels). Spheres were built around peak voxels at X = +39, Y = -49, Z = -26 and X = -36, Y = -43, Z = -26 for the FG (Figure A). And around peak voxels at X = -36, Y = -67, Z = -20 and X = +42, Y = -64, Z = -23 for the OFA (Figure B).

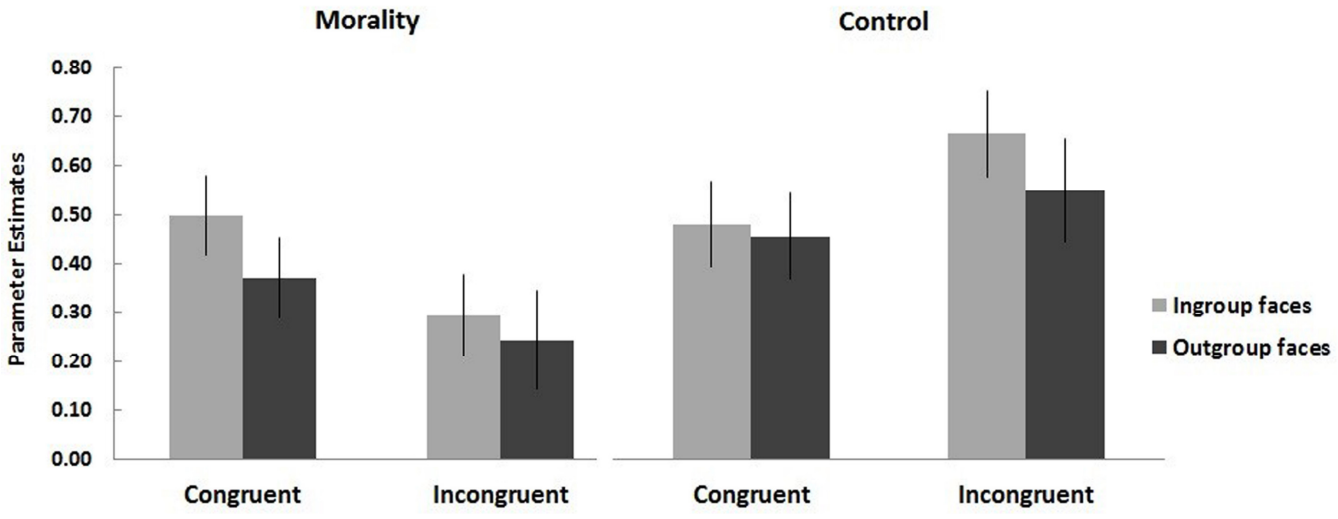

Figure 3.3. Within the ROI of the occipital face area (OFA) there was a significant interaction between target identity, congruency and task domain. Within the morality condition, activation was greater for viewing ingroup compared to outgroup faces and on congruent compared to incongruent trials.

Discussion

The goal of the current study was to examine the neural correlates of the motivation to behave in line with moral values. Complementing research on moral cognition –examining the cognitive processes involved in *thinking* about morality (i.e., moral reasoning and decision making) – we investigated the neural underpinnings of people's *behavior* on a task said the be indicative of their moral values. Specifically, we tested whether and how emphasizing the moral (compared to competence) implications of an implicit association test (IAT) would cause participants to inhibit their evaluative bias against Muslims. We used fMRI to study whether such an emphasis affects activation in brain areas implemented in visual attention towards facial stimuli and cognitive control – which would complement recent EEG research revealing enhanced perceptual attention and response-monitoring when the moral implications of an IAT are made salient (Van Nunspeet et al., 2014).

Our results revealed that visual attention towards the different facial stimuli in the IAT was dependent upon the emphasis on participants' morality: Participants in this condition showed greater activation in the occipital face area (OFA) when viewing ingroup compared to outgroup targets. Additionally, OFA activation was somewhat increased on congruent as compared to incongruent trials. These findings are consistent with the expectation that emphasizing the moral implications of the IAT affects participants' focus towards stimuli and perhaps their approach towards the task. As was shown by Van Nunspeet et al. (2014), participants who had read the moral test implications inhibited their responses on congruent trials. These are trials in which the easy (automatic) associations between Muslims and negative attributes and non-Muslims and positive attributes become evident.

To inhibit these prepotent responses, participants may need to be even more focused on the facial stimuli in the congruent compared to the incongruent trials to respond in line with their moral values. Furthermore, the fact that there was only a (marginally significant) effect of emphasizing moral concerns on initial (i.e., early) visual attention to faces and not on later and deeper facial processing in the fusiform gyrus, is in line with research showing that stressing moral test

implications is associated with increased social categorization on very early event-related brain potentials (i.e., the N1 and P150, occurring around 100 and 150ms after stimulus-onset; Van Nunspeet et al., 2014).

Table 3.1

Brain regions revealed by the Ingroup > Outgroup targets contrast.

Anatomical Region	L/R	voxels	Z	MNI coordinates		
				x	y	z
Fusiform Gyrus	R	3161	4.71	39	-49	-23
			4.70	36	-43	-23
			4.17	33	-61	-20
	L	3161	4.28	-36	-49	-20
			4.07	-33	-73	-11
			3.96	-36	-64	-14
(anterior) Medial Cingulate Cortex	R	486	4.42	9	5	34
			4.17	12	-7	52
			3.80	30	-25	46
Supramarginal Gyrus	L	160	4.76	-45	-1	10
			3.31	-33	8	19
			3.26	-57	2	7
Temporal Parietal Junction	L	141	4.39	-54	-25	22
			3.56	-42	-22	16

MNI coordinates for main effects, peak voxels reported with an FDR-corrected threshold of $p < .05$, with an extent threshold of 10 continuous voxels (voxels size was 3.0 x 3.0 x 3.0 mm).

Although visual attention was affected by the emphasis on the moral implications of the task, this did not affect participants' bias against Muslim women. This is different from previous studies (e.g., Van Nunspeet et al., 2014; Van Nunspeet et al., *under review*) in which participants who had read the moral test implications showed a smaller bias than participants who had read the implications concerning their competence. There it was argued that the emphasis on morality caused participants to inhibit their prepotent (automatic) prejudiced responses which resulted in the increased response times on congruent trials. However, compared to the study of Van Nunspeet et al. (2014) in which the interstimulus interval (ISI) lasted for just 500 milliseconds, the duration of the current ISI was

around two seconds. The inhibition of prepotent prejudiced responses may thus have occurred previous to stimulus onset or may have been undermined since participants had the time to prepare their response on the upcoming trial. In other words, the task could have become too easy to reveal implicit bias. Indeed, the amount of errors in the current research (4.5%) was only half of the error rates (8.3%) in research of Van Nunspeet et al. (2014). This explanation could also account for why we did not find greater activation in the neural regions associated with the regulation of implicit bias: The relatively easy IAT may have prevented participant from worrying about their performance. This is in line with research of Bengtsson, Lau, and Passingham (2009) who asked participants to perform either a significant (i.e., assessing their intelligence) or an insignificant (pilot test) experimental task. Their results revealed no differences in neural activation in prefrontal areas between the different types of tasks for correct responses. However, they did find that participants who performed the significant (compared to the insignificant) task showed increased neural activity on errors (Bengtsson et al., 2009). This is somewhat related to the study of van Nunspeet et al. (2014) in which it was shown that participants showed increased error-monitoring (i.e., greater error-related negativity modulations to *in*correct responses) when the moral (compared to the competence) implications of the IAT were emphasized. It is therefore possible that the difference between the motivation to perform in line with moral values as compared to one's competence is more evident on incorrect than correct responses. (Artificially) increasing the amount of errors during such an IAT and analyzing these events may thus reveal the differential neural activation we were aiming to find in the current research. Moreover, instructing participants to "clear their minds" when they see the fixation point may be crucial to overcome the effects of the increased ISIs (as was done in research by Beer et al., 2008). That is, to prevent participants to prepare their response on the upcoming trial.

Although we did not find an effect of the task instruction manipulation on the behavioral results, participants did show the typical IAT effect: A negative bias against Muslims. They responded more slowly on incongruent as compared to congruent IAT trials, indicating that associating outgroup members with positivity and ingroup members with negativity was more difficult for them than associating

outgroup members with negativity and ingroup members with positivity. However, the expected neural activation in regions associated with cognitive conflict and control –the ACC and DLPFC– was not evident for the incongruent > congruent contrast. It should be noted that not all fMRI studies that used an IAT have found these activation patterns. For example, Knutson et al. (2007) neither showed significant patterns of activation for incongruent compared to congruent trials when analyzing their *single* IATs (i.e., a gender and race IAT separately). Nevertheless, another experimental design (for example, in which IAT test blocks are presented repeatedly, alternating between blocks of congruent and incongruent trials) may have improved the BOLD response supposedly associated with the task demands.

Another aim of the current research was to extend previous behavioral and EEG research on the motivation to display moral behavior, by adding insights from an fMRI study revealing the particular brain areas involved in that motivation. Unfortunately, we were unable to expand current insights concerning increased cognitive control in case of an emphasis on moral concerns. This may have been due to the restrictions of the current research design mentioned previously (i.e., increased ISIs needed in an event-related fMRI experiment, and analyzing only correct responses since errors were too scarce), and illustrates the difficulty of optimizing an experimental paradigm for different scientific research methods (see also Scheepers, Ellemers, & Derks, 2013). On the other hand, we did find some additional support for increased visual attention towards targets when an IAT is presented as a measure of individual morality. And it may be the combination of such findings from different scientific research methods that can strengthen our knowledge of the underlying cognitive and neural mechanisms of moral motivation.

Conclusion

The current research revealed that when the moral implications of an IAT are emphasized, participants show greater activation in the occipital face area when they view pictures of ingroup compared to outgroup targets. Moreover, activation in this region was greater on (prejudice-) congruent compared to incongruent trials. In addition to previous research, these findings may suggest that especially people's

(visual) attention to a task increases once they have an opportunity to show their moral side.

Acknowledgements

We thank Gert-Jan Lelieveld for his help with the data collection and Jay van Bavel for helpful discussions about data analyses.

Part II

The importance of being perceived as moral by others

Chapter 4

Evaluation by an in- or outgroup member differentially affects moral task performance and the underlying cognitive mechanisms

This chapter is based on: Van Nunspeet, F., Derks, B., Ellemers, N., & Nieuwenhuis, S. Moral impression management: Evaluation by an ingroup member during a moral IAT enhances perceptual attention and conflict monitoring. *Manuscript under review*.

According to the Oxford dictionary being moral means "holding high principles for proper conduct". But what is considered 'proper'? Of course, individuals can have their own principles of what is good and bad. Nevertheless, the groups to which we belong (teams, organizations, or societies), and the group members to whom we feel connected, often define relevant standards of morality (see also Ellemers & Van den Bos, 2012). Behaving according to those standards is perceived as important: People are motivated to adjust their own behavior to moral (compared to competence) ingroup norms (Ellemers, Pagliaro, Barreto, & Leach, 2008), as a way to earn respect from fellow ingroup members (Pagliaro, Ellemers, & Barreto, 2011). Moreover, people identify more strongly with a moral than a competent group and are more proud to be a member of that group (Leach, Ellemers, & Barreto, 2007).

People's willingness to belong to moral groups and their pride in being a moral group member, can be explained by Social Identity Theory which proposes that people's self-views depend upon the groups to which they belong (Tajfel, 1978). Indeed, moral characteristics convey important social information: When asked to form an impression about other individuals, people are more inclined to gather information concerning morality than concerning competence or sociability (Brambilla, Rusconi, Sacchi, & Cherubini, 2011). Even when an impression has to be made within milliseconds, trustworthiness judgments are made faster than judgments of sociability and competence (Willis & Todorov, 2006). Moreover, people monitor their own behavior to maintain a moral self-image (Jordan & Monin, 2008). Due to the identity-defining function of morality –especially in group contexts, being moral is what we consider important in others and ourselves (Ellemers & Van den Bos, 2012).

The motivation to be moral elicits the tendency to adjust one's behavior to moral norms. This is not only evident in self-report measures (Pagliaro et al., 2011). For example, Van Nunspeet, Ellemers, Derks and Nieuwenhuis (2014) have shown that people adapt their implicit behavior when this is perceived as indicative of their morality: During an Implicit Association Test (IAT) participants were more inclined to control their negative bias towards Muslim women when they thought

the test measured their morality than when they thought the test measured their competence.

The reasoning that the significance of morality derives from its implications for people's social identity, leads to the prediction that the motivation to be moral should be particularly relevant in an ingroup context. Thus, we hypothesize that when participants are evaluated by an ingroup, rather than an outgroup member, they are more motivated to control their bias during performance on an IAT indicating morality (compared to competence).

Event-Related Brain Potentials and Moral Performance

The desire to be moral may elicit socially desirable answers pertaining to morality. This complicates the interpretation of self-reports on the importance of morality. Additionally, it remains unclear how people control their behavior to appear moral. Examining the cognitive processes involved in displaying moral behavior can elucidate how this is achieved.

In prior research, Van Nunspeet et al. (2014) revealed the cognitive processes that were associated with performance on a morally framed IAT. When test implications were presented in terms of morality compared to competence, participants' perceptual attention and response monitoring were enhanced during task performance. More specifically, event-related brain potentials (ERPs) suggested that participants paid more attention to the group membership of the photographed individuals presented in the IAT. This so-called social categorization was evident in modulations of the N1 and P150, two ERP components occurring around 100 and 200 ms after stimulus onset, that typically are larger when viewing ingroup vs. outgroup faces (e.g., Ito & Urland, 2003; Kubota & Ito, 2007). Van Nunspeet et al. (2014) argued that perceptual attention to the group membership of the women in the IAT was enhanced to enable participants to perform in line with their moral values.

Additionally, when morality instead of competence was emphasized in the IAT instruction, participants showed enhanced brain responses to the difference between incongruent and congruent trials and to errors. Specifically, the N450 and error-related negativity (ERN) modulations were larger when moral test implications were emphasized (Van Nunspeet et al., 2014). The N450, a negative

deflection around 400-500ms after stimulus-onset, is a component associated with conflict-monitoring, e.g. in language incongruencies (e.g., Nigam, Hoffman, & Simons, 1992), the Stroop task (e.g., Rebai, Bernard, & Lannou, 1997), and the IAT (Williams & Themanson, 2011). The ERN on the other hand, is a negative deflection within 100ms after a response is given. It is known to be larger for incorrect than correct responses (e.g., Gehring, Goss, Coles, Meyer, & Donchin, 1993; Nieuwenhuis, Blom, Band, & Kok, 2001), and for significant compared to non-significant errors (Hajcak, Moser, Yeung, & Simons, 2005). The findings of Van Nunspeet et al. (2014) thus suggest that the incongruency between the different IAT trials, as well as incorrect responses were perceived as more significant when the IAT was presented as a moral test. Additionally, their ERN results suggested that people are more concerned to show immoral than incompetent behavior.

Moral Performance in Group Contexts

In the current research we hypothesize that participants are more motivated to perform in line with moral values when they are being evaluated by a self-relevant other (an ingroup rather than an outgroup member). To examine this, we need to exclude alternative motivations to control bias, such as the wish to avoid offending the IAT target group in the presence of an ethnic outgroup member (in the current research, a Muslim woman; Lowery, Hardin, & Sinclair, 2001; Richeson & Ambady, 2003). This is why we introduced minimal categories: Based on a questionnaire ostensibly assessing personality styles, participants were evaluated by a non-Muslim individual who was presented as someone with the same (ingroup) or another personality type (outgroup).

We thus predict that participants will show a weaker IAT bias when the moral (compared to competence) test implications are emphasized, especially when they are evaluated by an ingroup (vs. outgroup) member. Extending the research of Van Nunspeet et al. (2014), we anticipate that participants who are evaluated by an ingroup member and to whom the moral test implications are emphasized will show increased perceptual attention towards pictures of Muslim versus non-Muslim women (indexed by N1 and/or P150 modulations) and enhanced conflict- and response-monitoring as indicated by the N450 and/or ERN. We tested these

hypotheses in two studies; an initial behavioral study (Study 4.1) and a follow-up study in which we recorded an electroencephalogram (EEG) during IAT performance (Study 4.2).

Study 4.1

Method

Participants and design.

Ninety-five non-Muslim students (3 males, M_{age} = 19.2 years, SD = 2.0) participated for money or course credits. One participant was excluded from the analyses, because s/he responded too late on more than 25% of trials, indicating lack of attention. Participants were randomly assigned to conditions in the 2 (task domain: morality/competence) X 2 (evaluator: ingroup/outgroup member) between-participants design.

Procedure.

After participants signed an informed consent in which it was explained that their participation could be recorded on video, they were seated in an individual room with a webcam, head phone and a camera placed in a top corner of the cubicle. Participants were told they would be paired with another participant based on questionnaire scores (ostensibly) assessing their personality styles and indicating whether they were either a so-called 'P'- or 'O'-type. After completing the questionnaire and a short pause, participants saw their own and other participants' scores (i.e., participant numbers were presented in combination with the 'P'- and 'O'-personality styles). Participants were then informed that they would cooperate either with a member of the same or a different group (as determined by their personality style). Then the IAT was introduced as a reaction time task during which the other person (i.e., the evaluator) would observe and give them feedback on every trial. After that, a webcam connection was simulated: The evaluator introduced him or herself and told that s/he would observe and provide feedback to the participant. A smile and thumbs up would follow a correct trial; frowning and pointing thumbs down an incorrect trial. Participants then either read about the moral or competent implications of the upcoming task, and started with the IAT.

In reality, all participants were said to have a 'P'-personality style and were introduced to a (same gender) confederate whose introduction was prerecorded. After the IAT, participants completed additional questions, were debriefed and thanked. The experiment lasted approximately fifty minutes.

Instruments.

The Implicit Association Test. Participants performed an IAT as designed by Greenwald, McGhee and Schwartz (1998). Stimuli representing the target concepts consisted of 10 pictures of non-Muslim and 10 pictures of Muslim women (faces without and with a headscarf respectively). Stimuli that represented positive and negative attributes consisted of 5 pictures of positive and 5 pictures of negative scenes, selected from the International Affective Picture System (IAPS; Lang et al., 2005).

For congruent trials, pictures of non-Muslim women shared the same response key as positive pictures and pictures of Muslim women the same response key as negative pictures. For incongruent trials, this was the case for non-Muslim women and negative pictures, and Muslim women and positive pictures. The order of the (in)congruent blocks was counterbalanced across participants. Training blocks 1, 2 and 4 consisted of 26 trials, test blocks 3 and 5 of 156 trials each. Every trial started with a fixation point, followed by a stimulus, a blank screen, and a feedback screen (see Figure 4.1). The feedback screen consisted of a movie clip (1250ms) of an evaluator showing either positive or negative feedback. To ensure that participants were aware that the evaluator was an in- or an outgroup member, two text displays indicated the group memberships of the participant and the evaluator. In case participants did not respond in time (i.e., within 680 ms), the feedback screen showed the words "too late".

Morality vs. competence task domain. Task domain was introduced using the instructions described in Van Nunspeet et al. (2014). Without mentioning the IAT design, or how performance would be measured, participants read the test would indicate their moral values concerning egalitarianism in the morality condition, or their ability to learn new tasks in the competence condition. In both conditions, participants were instructed to respond as quickly and accurately as possible. The test implications were repeated before each test block.

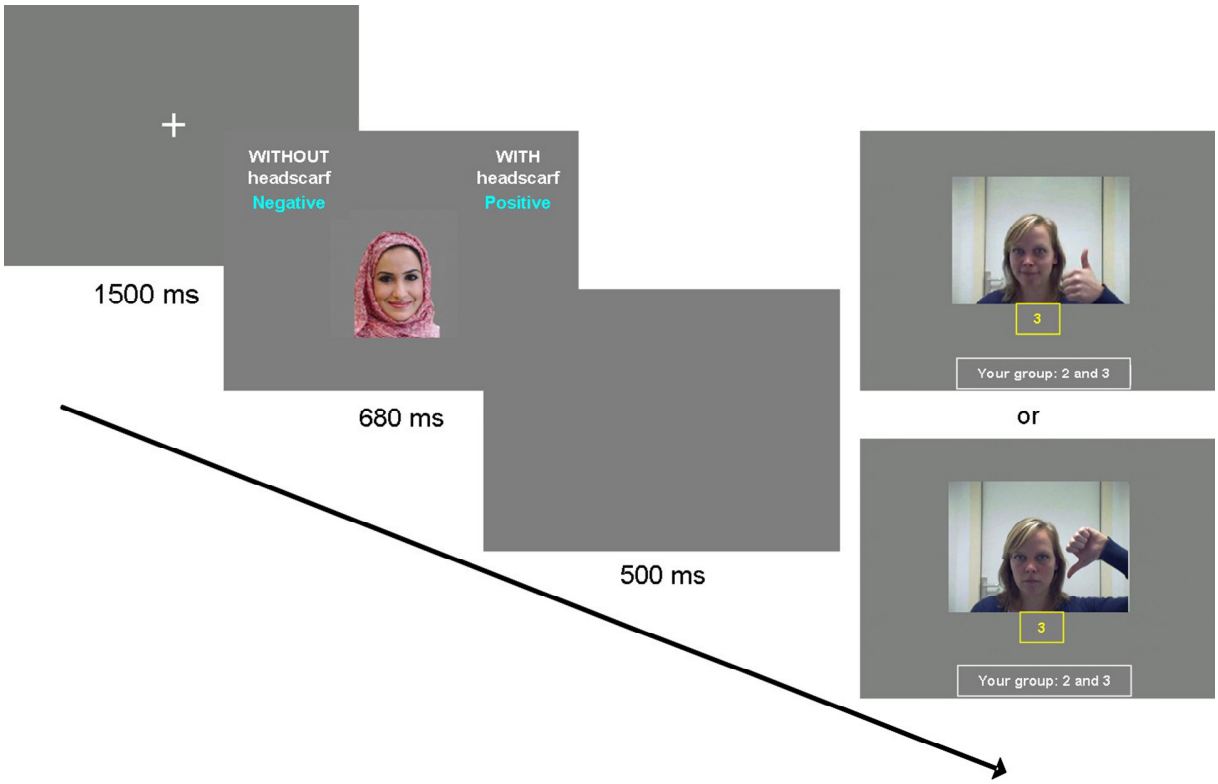

Figure 4.1. An IAT trial. The feedback screen was a movie display (1250ms) in which the confederate (here displayed as ingroup member) gave positive or negative feedback.

The IAT effect. The dependent measure was the IAT effect (i.e., the *D* score), which was calculated as the difference in reaction times on incongruent and congruent trials divided by a pooled *SD* of all correct trials (Greenwald et al., 2003; see also Van Nunspeet et al., 2014).

Checks. To check that the perceived validity of the IAT did not differ between the conditions, we asked participants to respond to the statement: "My test score can assess what kind of person I am". Furthermore, we asked to what extent participants hoped to have made a good impression on the evaluator: "I hope the evaluator has the impression that I am competent/kind/moral" (3 items, $\alpha = .90$). Identification with the P-type group was checked with two items ("I identify strongly with the P group" and "I feel equal to the other group members in terms of general attitudes and beliefs"; $r = .41$, $p < .001$). Participants could respond on a 7-point scale (1: "completely disagree" to 7: "completely agree").

Results and Discussion

Checks.

As intended, participants in the four experimental conditions did not differ in their ability to identify with the experimentally created ingroup (grand-average M = 3.77, SD = 1.20); $F(3, 90) = 1.37$, $p = .26$, and did not think differently about the perceived validity of the test; M = 3.64, SD = 1.62; $F(3,90)<1$. In line with prior findings, participants in the morality condition indicated positive impression management to be more important than participants in the competence condition; $M_{morality} = 4.83$, $SD = 1.01$; $M_{competence} = 4.28$, $SD = 1.04$; $F(1,90) = 6.58$, $p = .01$, $\eta^2 = .07$. There was no effect of evaluator nor an interaction effect; F's < 1.49, p's > .23, indicating the importance of the moral task was enhanced, independently of whether participants were evaluated by an in- or an outgroup member.

IAT effect.

Overall, participants showed the standard IAT effect, indicating a negative implicit bias towards Muslim women; $t(93) = 6.83$, $p < .001$. More errors were made on incongruent than on congruent trials; respectively M = 9.35, SD = 7.01 and M = 6.46, SD = 5.40; $t(93) = 4.50$, $p < .001$; this was not affected by task domain or evaluator , all F's < 1.87, p's > .18. Consistent with previous research (Van Nunspeet et al., 2014), an ANOVA with the D score as dependent variable and domain and evaluator as independent factors, revealed a significant main effect of domain; $F(1,90) = 5.57$, $p = .02$, $\eta^2 = 0.06$. Overall, participants in the morality condition showed a smaller IAT effect than participants in the competence condition, M(morality) = 0.18, SD = 0.34; M(competence) = 0.36, SD = 0.39[4]. Additionally, we found the predicted interaction effect between domain and evaluator; $F(1,90) = 4.26$, $p = .04$, $\eta^2 = 0.05$ (see Figure 4.2), indicating that participants who were evaluated by an ingroup member showed significantly less bias in the morality than in the competence condition; M(morality) = 0.10, SD = 0.32; M(competence) = 0.43, SD = 0.33; $F(1,90) = 9.82$, $p < .01$, $\eta^2 = .10$, while

[4] Consistent with previous research (Van Nunspeet et al., 2014), this difference was related to increased response latencies on congruent trials in the morality compared to the competence condition; M(morality) = 494.85, SD = 20.10; M(competence) = 480.65, SD = 16.61; $F(1,93) = 13.62$, $p < .001$, $\eta^2 = .13$.

this was not the case when evaluated by an outgroup member; *M*(morality) = 0.26, *SD* = 0.34; *M*(competence) = 0.29, *SD* = 0.44; *F* < 1. These findings extend previous research by showing that moral impression management is particularly important in an intragroup context (even if the broader significance of the ingroup is relatively minimal). In Study 4.2 we examine what cognitive processes are associated with the tendency to conform to moral values in group contexts.

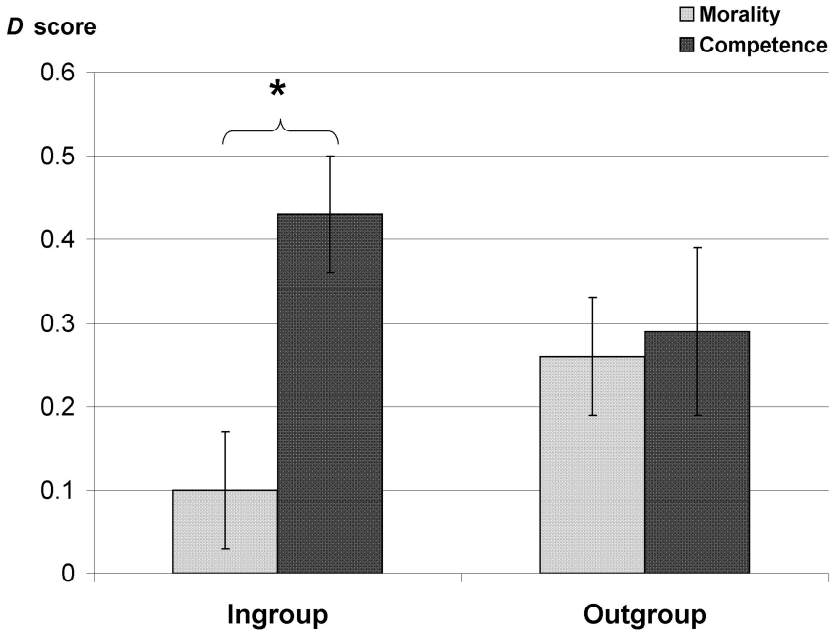

Figure 4.2. Condition means relevant to the interaction effect on the *D* scores.

Study 4.2

Method

Participants and design.

Sixty-seven non-Muslim, right-handed, healthy students (18 males, M_{age} = 20.6 years, *SD* = 2.1) participated for money or course credits. Three participants were excluded from all analyses because they responded too late on more than 25% of the trials, indicating lack of attention. Two other participants could not be included in the analysis of self-report data because they failed to complete the questions, and

four participants had to be excluded from the ERP analyses, because of technical problems during the EEG acquisition. Remaining participants were randomly distributed across conditions of the 2 (domain: morality/competence) X 2 (evaluator: ingroup/outgroup member) between-participants design.

Procedure.

The procedure and measures were similar to Study 4.1, with the following exceptions. Participants completed the questionnaire to ostensibly determine personality style before they came to the EEG lab. The feedback screens in the IAT consisted of a photograph of the confederate instead of a movie display. Finally, to elicit a sufficient number of errors to reliably estimate the ERN, the maximum duration of the stimulus presentation was reduced from 680ms to 550ms, and the total number of test trials increased to 600 (300 congruent and 300 incongruent trials).

EEG acquisition.

The EEG was recorded from 19 Ag/AgCl scalp electrodes mounted in an elastic cap, and from the left and right mastoids, using a 19-channel Biosemi active-electrode recording system (sampling rate 256 Hz). To assess horizontal and vertical eye movements, electrodes were placed on the outer canthi of the eyes and approximately 1 cm above and below the participant's right eye. EEG activity was recorded using ActiView software, offline data analyses were performed using Brain Vision Analyzer, and the experiment was presented with E-prime software. The EEG signal was referenced off-line to the average mastoid signal, corrected for ocular and eye-blink artifacts using the method of Gratton, Coles, and Donchin (1983), and filtered (1-15 Hz). Single-trial stimulus-locked and response-locked epochs were extracted, ranging from -300ms to 1000ms after the event. These epochs were subjected to artifact rejection, then averaged and baseline-corrected by subtracting the average signal value between 200-0ms pre-stimulus or between 300-50ms prior to the response. Separate stimulus-locked ERP epochs were created for correct congruent and incongruent trials with pictures of Muslim and non-Muslim women. Separate response-locked ERP epochs were created for correct and error trials.

ERP analyses.

Visual inspection of the data indicated that the N1, P150, and ERN components were most evident at midline electrode sites FCz and Cz. The N450 was most evident at CPz and Pz. The stimulus-locked ERP components were quantified as the peak amplitude within a time window post-stimulus (N1: 90-110ms; P150: 100-250ms; N450: 325-500ms), whereas the ERN was quantified as the peak amplitude of the signal between -50 and 150ms around the response. Each average ERN was based on at least 10 trials[5]. Peak amplitude values of the N1, P150, and N450 were submitted to a 2 (electrode site: FCz/Cz or CPz/Pz [N450]) x 2 (target: Muslim/non-Muslim women) x 2 (congruency: congruent/incongruent) mixed-model ANOVA. Peak amplitude values of the ERN were submitted to a 2 (electrode site) x 2 (accuracy: correct/error) x 2 (congruency) mixed-model ANOVA. In every analysis, domain (morality/competence) and evaluator (ingroup/outgroup) were included as between-participants factors[6].

Results and Discussion

Behavioral results.

Checks. As intended, identification with the ingroup (2 items, $r = .50$, $p < .001$) was equal across experimental conditions (grand-average $M = 3.53$, $SD = 1.36$), $F(1, 58) < 1$, as was the perceived validity of the test; $M = 3.58$, $SD = 1.56$; $F(1,58) < 1$. Again, participants in the morality condition indicated more concern about impression management than in the competence condition; $M_{morality} = 5.25$, $SD = 0.83$; $M_{competence} = 4.63$, $SD = 0.82$; $F(1,58) = 8.39$, $p = .01$, $\eta^2 = .13$.

IAT effect. Overall, participants showed the standard IAT effect, indicating a negative implicit bias towards Muslim women; $t(63) = 5.46$, $p < .001$. IAT effects were not systematically affected by evaluator or task domain; F's < 1, p's $> .1$, indicating that the emphasis on morality or competence and the group membership of the evaluator was not visible in task performance. This likely is due to the changes we made to optimize the task for ERP recordings: To ensure enough

[5] Some participants made less than 10 errors, explaining different degrees of freedom between the stimulus- and response-locked ERP analyses.

[6] Electrode site was not of interest for the current research, see Appendix B for significant interaction effects with this factor.

errors to reliably estimate the ERN, the maximum response time was reduced. In Study 4.1 –and in previous research (Van Nunspeet et al., 2014)– participants controlled their bias by delaying responses on congruent trials, which may have been impossible in this study, given the tight response deadline. A follow-up study corroborates this explanation. When we examined behavioral effects of task instruction and ingroup/outgroup evaluators using a response window of 680 ms (as in Study 4.1 and prior research), the IAT bias was significantly lower in the morality compared to the competence condition, when participants were evaluated by a minimal ingroup member (Van Nunspeet, Ellemers, & Derks, *manuscript under review*).

Nonetheless, the identity of the evaluator did affect behavioral responses in the current data. Besides the fact that more errors were made on incongruent (M = 34.4, SD = 18.4) than congruent trials (M = 25.6, SD = 15.5); $t(63) = 4.87$, $p < .001$, participants in the ingroup evaluator condition made fewer errors (M = 50.3, SD = 24.9) than participants in the outgroup evaluator condition (M = 70.4, SD = 33.4); $F(1,60) = 7.28$, $p = .01$, $\eta^2 = .11$. This is consistent with our reasoning that participants are generally more motivated to perform well when evaluated by an ingroup member.

ERP results.

Perceptual attention.

N1. The N1 results revealed the expected evidence of categorization: The N1 was larger for pictures of Muslim women (M = -7.18 μV, SE = 0.35) than non-Muslim women (M = -6.91 μV, SE = 0.35); $F(1,56) = 3.52$, $p = .07$, $\eta^2 = .06$ (see Figure 4.3). The predicted interaction between target, domain and evaluator was significant; $F(1,56) = 4.36$, $p = .04$, $\eta^2 = .07$. Separate analyses for ingroup vs. outgroup evaluators revealed a marginally significant interaction between target and task domain in case of an ingroup evaluator; $F(1,29) = 3.53$, $p = .07$, $\eta^2 = .11$, but not in case of an outgroup evaluator; $F(1,27) = 1.02$, $p = .32$. Separate analyses per task domain revealed a significant target by evaluator interaction in the moral domain; $F(1,31) = 6.69$, $p = .02$, $\eta^2 = .18$, but not in the competence domain; $F<1$. As a result, categorization was significantly enhanced in the morality/ingroup

condition (F[1,56] = 11.35, p = .001, η^2 = .17), but not in the other conditions (F's<1; see Figure 4.4).

P150. Analyses of the P150 only revealed the expected main effect of target: The P150 was larger for pictures of Muslim women (M = 5.44 μV, SE = 0.48) than non-Muslim women (M = 3.77 μV, SE = 0.43); F(1,56) = 93.13, p < .001, η^2 = .62 (see Figure 4.3).

This suggests that enhanced social categorization of (non-)Muslim women in case of moral task performance under ingroup evaluation, only occurs in initial stages of perceptual attention (N1).

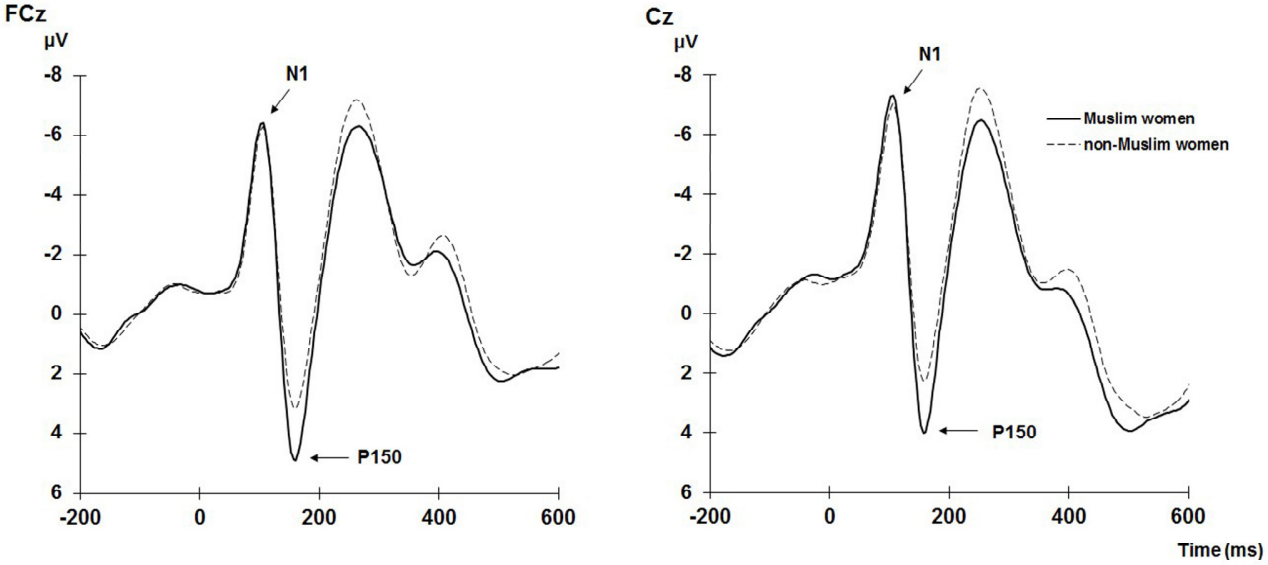

Figure 4.3. Differences in N1 and P150 amplitudes for pictures of Muslim and non-Muslim women. Only the N1modulation interacted with task domain and evaluator.

Conflict- and response-monitoring.

*N450.*Results showed the anticipated effect of congruency: The N450 was larger for incongruent (M = -0.13 μV, SE = 0.33) compared to congruent (M = 0.64 μV, SE = 0.40) trials; F(1,56) = 5.92, p = .02, η^2 = .10 (see Figure 4.5).

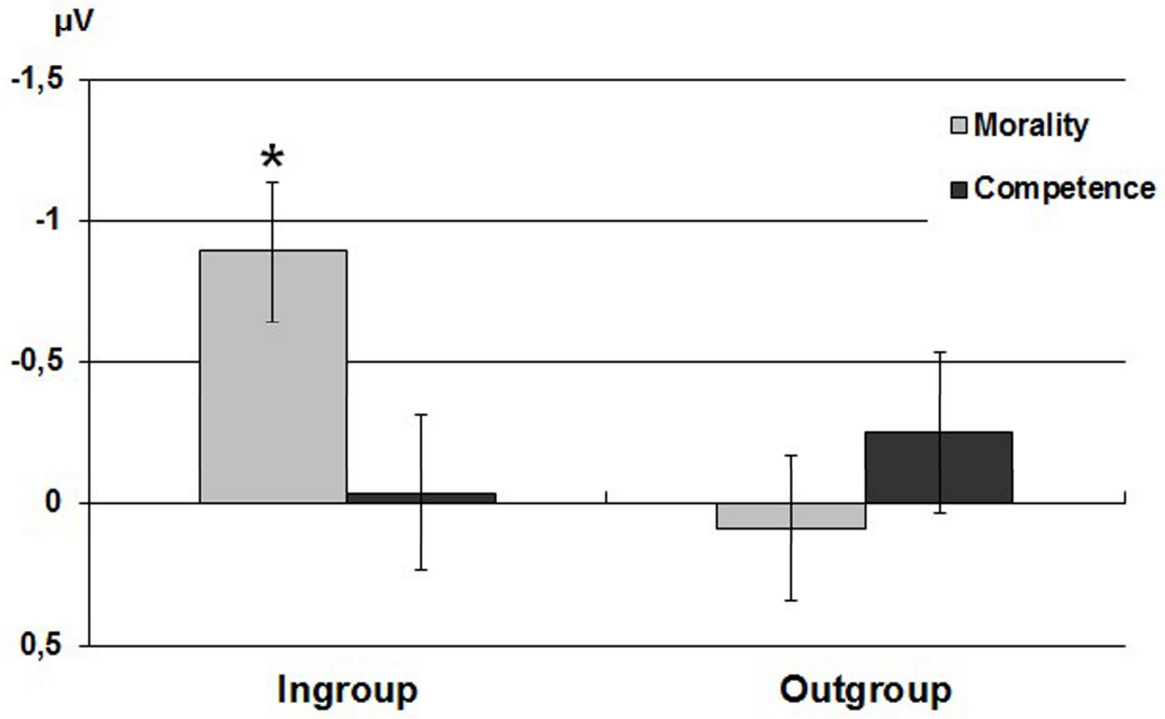

Figure 4.4. The mean differences in N1 amplitude between Muslim vs. non-Muslim targets for each condition.

There was also a main effect of target: The N450 was larger for non-Muslim (M = -0.28 μV, SE = 0.37) compared to Muslim women (M = 0.79 μV, SE = 0.33); $F(1,56) = 24.06$, $p < .001$, $\eta^2 = .30$. Importantly, both main effects were qualified by a significant four-way interaction between congruency, target, domain and evaluator; $F(1,56) = 5.75$, $p = .02$, $\eta^2 = .09$. Separate analyses for the task domain conditions revealed a significant interaction between congruency, target and evaluator in the morality condition; $F(1,31) = 5.36$, $p < .03$, $\eta^2 = .15$, but not in the competence condition; $F(1,25) = 1.30$, $p = .27$.

Furthermore, in the morality condition, there was an interaction between congruency and target in the ingroup evaluator condition; $F(1,16) = 10.26$, $p = .006$, $\eta^2 = .39$, but not in the outgroup evaluator condition; $F<1$. The N450 modulation on incongruent compared to congruent trials in the morality/ingroup condition was significant when viewing pictures of non-Muslim women; $F(1,16) =$

6.45, p = .02, η^2 = .29, and not when viewing Muslim women; F<1 [7,8] (see Figure 4.6). These results suggests that conflict-monitoring was enhanced (on non-Muslim trials) when moral test implications were stressed and participants were evaluated by an ingroup member.

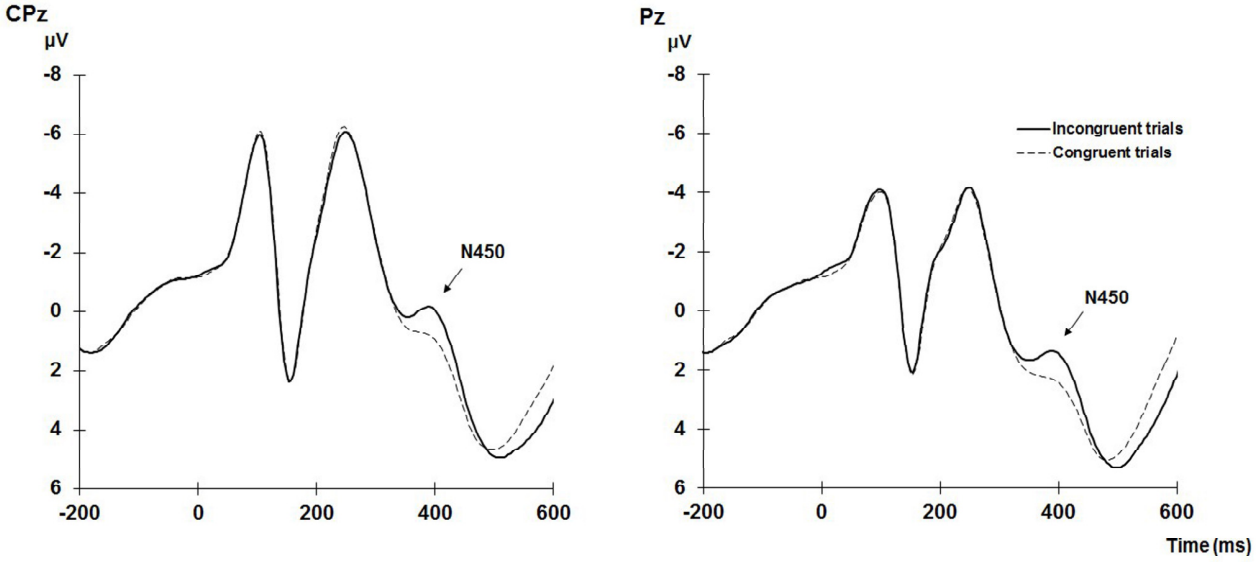

Figure 4.5. Differences in N450 amplitudes for incongruent and congruent trials.

ERN. As anticipated, results showed that the ERN was larger for incorrect (M = -6.90 µV, SE = 0.69) than correct trials (M = 2.95 µV, SE = 0.46); F(1,44) = 173.52, p < .001, η^2 = .80. There was a marginally significant interaction effect between accuracy and task domain; F(1,44) = 3.37, p = .07, η^2 = .07, indicating that the ERN modulation was somewhat larger in the competence ($M_{difference}$ = -1.22 µV, SE = 1.12; F[1,44] = 100.07, p < .001, η^2 = .70) than the morality condition ($M_{difference}$ = -8.48 µV, SE = 0.99; F[1,44] = 73.49, p < .001, η^2 = .63). More importantly however, the ERN modulation in the morality and competence

[7] Findings of N450 modulations for targets can be found in Appendix B.
[8] This may reflect the specific nature of our paradigm, in which these trials confronted participants with pictures of a non-Muslim target, while receiving feedback from a non-Muslim evaluator, arguably increasing the need for conflict-monitoring.

conditions differed depending on evaluator type: There was a marginally significant between-subjects interaction effect of task domain and evaluator; $F(1,44) = 3.59$, $p = .07$, $\eta^2 = .08$[9]. Even though the simple contrasts were not significant (F's < 2.32, p's > .14), the means pattern indicates a reversal of the effect. Response monitoring was enhanced under ingroup evaluation in the morality (M = -2.67 μV, SE = 0.83) compared to the competence condition (M = -1.36 μV, SE = 0.94), but enhanced in the competence (M = -2.96 μV, SE = 0.98) compared to the morality condition (M = -0.88 μV, SE = 0.86) under outgroup evaluation (see Figure 4.7).

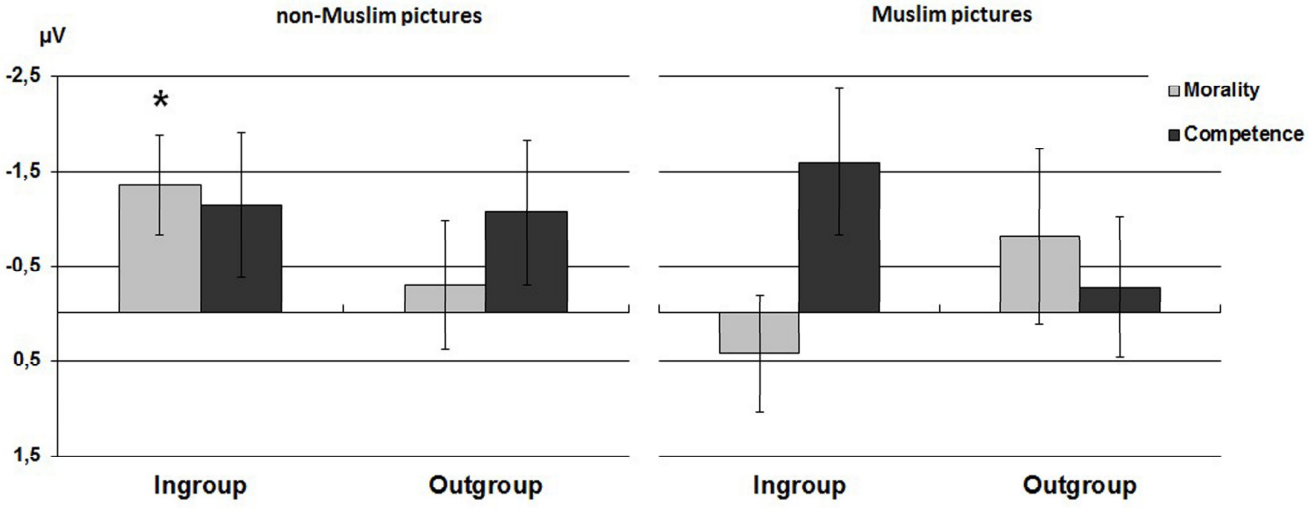

Figure 4.6. The mean differences in N450 amplitude between incongruent vs. congruent trials for each condition.

[9] To clarify this marginally significant effect, we conducted separate analyses for FCz and Cz. Results showed that both interaction effects were only significant at Cz (accuracy*domain: $F[1,44] = 3.98$, $p = .05$, $\eta^2 = .08$; domain*evaluator: $F[1,44] = 4.07$, $p = .05$, $\eta^2 = .09$).

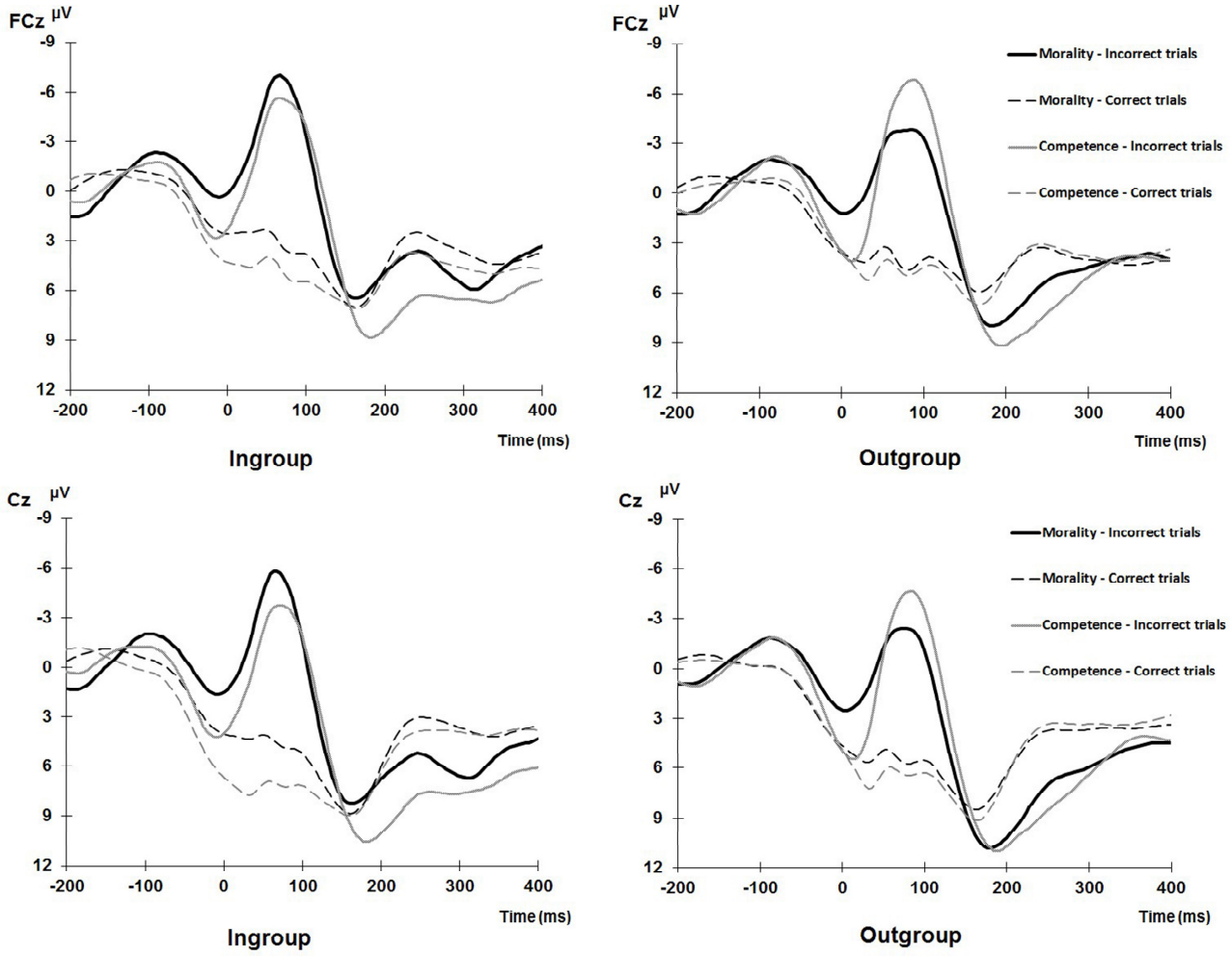

Figure 4.7. The error-related negativity in the morality and competence conditions in case of an ingroup (left) or outgroup evaluator (right).

General Discussion

The current studies extend previous research on the motivation to comply with moral ingroup norms (Ellemers et al., 2008; Pagliaro et al., 2011). We discovered that participants controlled implicit bias when the moral implications of an IAT were emphasized and when they were evaluated by a (self-relevant) ingroup member. In Study 4.1, participants responded more slowly on congruent IAT trials, suggesting inhibition of prepotent reaction tendencies possibly revealing prejudice. Complementing prior research (Van Nunspeet et al., 2014), ERP results in Study 4.2 revealed that this was associated with enhanced perceptual attention and social categorization of the target women in the IAT (as indicated by the N1). Participants were thus more focused on the identity of the different targets presented, which is needed to control biased responses. Thus, emphasizing the moral implications of the IAT does not make people insensitive to social categorizations. Instead, it triggers increased perceptual attention in order to adjust behavior. Indeed, previous ERP research has revealed that similar early attentional processes can be moderated by motivational states (e.g., Amodio, 2010; Cunningham, Van Bavel, Arbuckle, Packer, & Waggoner, 2012). Thus, our findings help understand how people control their prejudice towards Muslim women, to show they are moral in front of self-relevant others. Note however that this is different from the attempts to appear unprejudiced towards target group representatives, as revealed by Lowery et al. (2001) and Richeson and Ambady (2003). Complementing this prior work, we reveal that bias control can also be affected by the importance of sharing moral norms with one's ingroup.

Conflict- and response monitoring (indicated by the N450 and ERN) were also affected by the moral or competence implications of the IAT and the ingroup vs. outgroup evaluator. That is, the detection of incongruent compared to congruent trials (N450 modulation) was enhanced when participants in the morality/ingroup condition viewed non-Muslim women. Moreover, whereas response monitoring (ERN on correct and incorrect trials) seemed to be enhanced in the morality compared to the competence condition when the evaluator was an ingroup member, this pattern was reversed when the evaluator was an outgroup member. Since ERN amplitudes have been found to be related to estimates of

control (Amodio, et al., 2004), this suggests increased motivation to control bias towards Muslim women in the moral ingroup condition. We did not anticipate participants to be particularly sensitive to competence task instructions when evaluated by an outgroup member. However, a similar (non-significant) reversal of the importance of competence vs. morality depending on the group membership of the evaluator was observed in the behavioral results of Study 4.1 and the N1 results in Study 4.2: Whereas behavioral bias on the moral IAT was reduced in the ingroup evaluation condition, bias on the competence IAT was diminished in the outgroup evaluation condition. Likewise, the N1 modulation was greater in the morality/ingroup than in the morality/outgroup condition, while it was somewhat larger in the competence/outgroup than the competence/ingroup condition. Although (probably due to limited statistical power) these effects did not reach significance, they could suggest that whereas moral impression management is more important in the ingroup, displaying competence is more relevant towards the outgroup. Future research could further examine this. The current findings demonstrate the importance of morality for self and social identity, by revealing that people are especially motivated to adjust their moral task performance when monitored by a self-relevant group; this is associated with increased perceptual attention and conflict monitoring.

Acknowledgements

We thank Tamar van Herk, Laura de Reus, Ilona Domen, and Alma Vermeulen for their help with data collection; Marieke Visser and Wouter Steijn for their contribution to the experiment; and David Amodio for his advice concerning data analyses.

Chapter 5

Controlling implicit prejudice: The effects of moral implications, and evaluation by (non)significant others

This chapter is based on: Van Nunspeet, F., Ellemers, N., & Derks, B. Reducing implicit prejudice against Muslim women: The effects of moral concerns, intra- and intergroup motives. *Manuscript under review.*

The study of attitudes, stereotypes and prejudice is often complicated by social desirability issues: People sometimes adjust their explicit attitudes to appear unbiased (e.g., Crosby, Bromley, & Saxe, 1980). The development of implicit measures of prejudice that capture more automatic biases against social (out)groups was seen to offer a solution to this problem. People may display implicit biases even while they explicitly endorse egalitarian views (e.g., Dovidio, Kawakami, & Beach, 2001), and this is why it is often suggested that implicit prejudice captures the 'automatic' evaluative associations with other groups.

A popular and widely used implicit measure of prejudice is the Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998). The IAT is based on the idea that it is easier to associate your ingroup with positive attributes and an outgroup with negative attributes than vice versa. As a result, participants tend to respond faster on trials in which pictures of ingroup members are associated with positive stimuli (using the same response key) and outgroup members with negative stimuli (congruency). By comparison, they respond more slowly on trials in which ingroup members are associated with negative stimuli and outgroup members with positive stimuli (incongruency). The difference between response latencies on incongruent and congruent trials is taken to assess the degree of implicit bias against a social outgroup.

Although the IAT is frequently presented as a measure of automatic bias, by now several studies have shown the malleability of 'automatic prejudice. This suggests that implicit biases can be influenced too, for example by self-concerns and social motives (for an overview see Blair, 2002). Effects of self-concerns are shown in research were the induction of stereotype threat among Whites –by triggering the stereotype that they are racists– increased implicit biases towards Blacks (Frantz, Cuddy, Burnett, Ray, & Hart, 2004; Rudman, Dohn, & Fairchild, 2007). Other research has revealed that implicit biases can also be affected by intergroup concerns: When a Black experimenter was present during participants' performance on an IAT, Whites were able to inhibit their pro-White bias (e.g., Lowery, Hardin, & Sinclair, 2001). Additionally, Richeson and Ambady (2003) showed the significant effect of the role of such a Black person present: Their participants also displayed a smaller bias against Blacks, but only when their Black

partner in the experiment was a superior (instead of a subordinate). Furthermore, Van Nunspeet, Ellemers, Derks, and Nieuwenhuis (2014) revealed that emphasizing the moral implications of performance on an IAT –compared to implications concerning individual competence– led participants to show a smaller bias against Muslims. In a follow-up study, this effect was particularly strong when people were evaluated by minimal ingroup (rather than outgroup) members, thus suggesting effects of intragroup concerns (Van Nunspeet, Derks, Ellemers, & Nieuwenhuis, under review).

Current research

Although different motives and contexts have been shown to affect people's evaluative bias, to our knowledge they have not been directly compared in one study. It is thus unclear which concern or motive would benefit the control of bias against an outgroup when for example, interpersonal contact with a person from the target group is not feasible. In the current research, our aim is to examine the effects of three different interventions on people's ability to control their evaluative bias against an outgroup in one IAT experiment: (1) Personal concerns about moral implications of displaying bias; (2) intergroup motives (i.e., concerns about displaying bias in front of a representative of the devalued group) and (3) intragroup concerns about displaying bias in front of self-relevant others. Specifically, we demonstrate how people's evaluative bias against Muslims is affected by (1) emphasizing the moral (compared to competence) test implications of the IAT; (2) having participants be observed by either a Muslim or a non-Muslim evaluator (first ingroup/outgroup dimension); and (3) presenting this evaluator as either a minimal ingroup or outgroup member (second ingroup/outgroup dimension, resulting in cross-categorization). In the current study we combined these interventions to directly compare their effects on reducing implicit evaluative bias and to examine whether and how they may influence one another.

Additionally, we aimed to examine the underlying processes associated with reducing implicit bias. In studies concerning the effects of personal and social motives on people's evaluative biases, little attention has been devoted to how such a bias (i.e., IAT performance) was affected. In an IAT, bias is reduced by

diminishing the difference between response latencies on stereotype-incongruent and stereotype-congruent trials. However, this can be accomplished in two ways: Either by becoming quicker on incongruent trials (and thus becoming better in associating the outgroup with positive attributes), or by responding more slowly on congruent trials (and inhibiting negative associations with the outgroup and positive associations with the ingroup). Interestingly, the smaller IAT effect in research of Richeson and Ambady (2003) was due to slower responses on congruent trials. In a similar vein, Van Nunspeet et al. (2014) revealed that an emphasis on morality caused participants to show a smaller IAT bias, caused by their slowed down responses on congruent trials. In addition, these researchers showed that stressing the moral test implications was associated with enhanced response-monitoring (measured using EEG). results suggested that participants' reduced bias was related to the inhibition of prepotent responses on stereotype-consistent (i.e., congruent) trials (Van Nunspeet et al., 2014). In the current research, we therefore examined the pattern of response latencies on congruent and incongruent trials separately to see how exactly the three types of interventions affected participants' evaluative bias.

Study 5.1

Method

Participants.

Only female, non-Muslim, students (N = 225; M_{age} = 20.5 years, SD = 2.6) participated in the study and received either money or course credit for their participation. Two participants were excluded from analyses: One due to technical problems, another because she responded too late on all IAT trials, indicating lack of attention. Participants were randomly assigned to one of the eight experimental conditions of the 2 (Task Domain: morality/competence) x 2 (Evaluator's Minimal Group: ingroup/outgroup) x 2 (Evaluator's Religion: Muslim/non-Muslim) between-participants design. Note that the evaluator was the same individual in all conditions, but that she did or did not wear a headscarf (see Figure 5.1).

Procedure.

Participants were seated in an individual computer room with a webcam on top of the computer screen, and a camera behind them in a top corner of the

cubicle. They were told that they would be working together with another participant. They then completed a (bogus) questionnaire that was said to assess whether they had either a so-called 'P'- or 'O'- personality style. After a short waiting period, participants learned about their own alleged personality style and the styles of the other participants and they were informed whom they would be working with during the experiment. The other person either was said to have the same personality style as the participant (to convey this individual was a member of the same minimal group as the participant), or she allegedly had the other personality style (to indicate this individual belonged to a different group). Participants then read that they would perform a computer task. During the first part of the experiment, the other person would supposedly observe and give them feedback after every trial and the roles would be reversed in the second part. Thereafter, a webcam connection was simulated: The other person introduced herself and said that she would observe and provide visual feedback on every trial. Then, participants read either the morality or competence instruction and started with the IAT. In reality, all participants were said to have a 'P'- personality style and were introduced to a confederate whose movies were prerecorded. Feedback displays during the IAT were related to participants' actual responses (i.e., positive feedback when they responded correctly, negative feedback when they responded incorrectly). After the IAT, participants completed some self-report items and were properly debriefed.

Task domain manipulation. Before the start of the IAT, half of the participants read that the computer task they were going to perform could indicate their endorsement of *moral values* concerning egalitarianism and discrimination (the morality condition). The other half of the participants was informed that the test could indicate their *ability* to process new information and to learn new tasks (the competence condition). All participants were instructed to respond as quickly and accurately as possible and the test implications were repeated before the start of each test block (see also Van Nunspeet et al., *2014*).

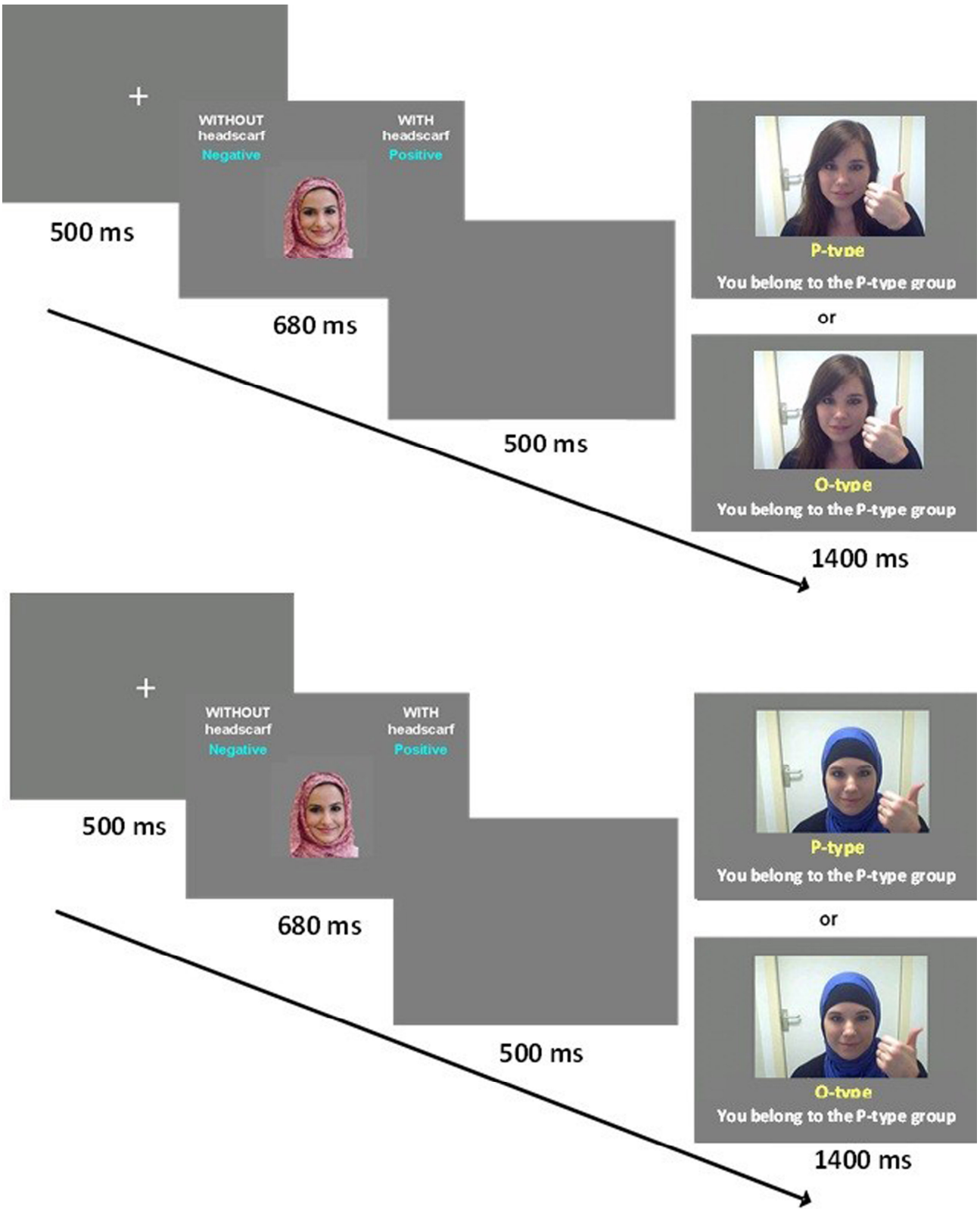

Figure 5.1. Example of an (incongruent) IAT trial. The (same) evaluator resembled either a non-Muslim (top) or Muslim (bottom) woman.

Instruments.

The Implicit Association Test. Participants performed the five blocks of the IAT as designed by Greenwald et al. (1998). Stimuli representing the target concepts consisted of 10 pictures of Muslim women (wearing a headscarf) and 10 pictures of non-Muslim women (not wearing a headscarf). Stimuli that represented positive and negative attributes consisted of 5 pictures of positive scenes, and 5 pictures of negative scenes, selected from the International Affective Picture System (Lang et al., 2005).

In (training) block 1, participants were asked to respond to the pictures of women by pressing a left key for Muslim women and a right key for non-Muslim women. In (training) block 2 they were asked to use the same two keys to respond to the negative and positive pictures. In block 3 (a test block) both picture types were presented and participants responded with one key to pictures of both Muslim women and negative scenes and with the other key to pictures of both non-Muslim women and positive scenes (i.e., congruent trials). In (training) block 4, the response keys for the pictures of (non-)Muslim women were switched and in block 5 (a test block), participants had to respond to pictures of both non-Muslim women and negative scenes with one key and to pictures of both Muslim-women and positive scenes with one other key (i.e., incongruent trials). Blocks 1, 2 and 4 consisted of 20 trials, blocks 3 and 5 of 70 trials each. Every trial started with a fixation point (500 ms), followed by stimulus presentation (680 ms), a blank screen (500 ms) and a feedback screen (1400 ms). The feedback screen consisted of a movie clip of the evaluator showing either positive (smiling and holding 'thumbs up') or negative (frowning and pointing 'thumbs down') feedback. To ensure that participants were aware of the minimal group membership of their evaluator, we inserted a text display below the movie indicating the personality type of the evaluator, and a text display at the bottom of the screen indicating the personality type group of the participant (see Figure 5.1). In case participants did not respond in time, they saw the words "too late".

The IAT effect. The dependent measure was the IAT effect, indicated by the *D* score, and measured as the difference in reaction times on incongruent and congruent trials divided by a pooled *SD* of all correct trials (according to the

scoring algorithm described by Greenwald et al. 2003). We included all trials, replaced error latencies with a replacement value (M + 2 $SD_{correct}$) and replaced latencies exceeding the maximum response time with the maximum response time of 680 ms. The resulting positive D scores are an indication of people's evaluative bias against Muslim women.

Checks. Directly after the IAT, we checked the task domain manipulation: Participants were asked to indicate what the IAT intended to measure. They could indicate that the test either measured how well they were able to process information and to learn new tasks, or that it assessed their moral values concerning egalitarianism and discrimination. Second, we checked the evaluator's minimal group manipulation by asking participants to indicate whether their evaluator was a member of the same or another minimal group. Furthermore, we tested participants' perceptions of the validity of the test (i.e., "My test score can assess what kind of person I am"), and their overall impression of their evaluator ("I think the participant who gave me feedback is competent/kind/moral", 3 items). Participants could respond on a 7-point Likert scale (1 = completely disagree, 7 = completely agree).

Results

Checks.

Results concerning the manipulation of task domain showed that 96% (N = 105) of participants in the morality condition indicated that the test measured their moral values concerning egalitarianism and discrimination. Moreover, ninety-seven percent (N = 110) of participants in the competence condition indicated that the test measured their ability to quickly process information and learn new tasks. Results concerning the evaluator's minimal group manipulation showed that 95% (N = 103) of participants whose evaluator was an ingroup member correctly answered that their evaluator was a member of their own group. One hundred percent (N = 115) of participants whose evaluator was an outgroup member answered correctly that their evaluator was a member or the other group. Excluding the participants who answered one of the checks incorrectly (N = 10) did not alter the pattern of the means. We therefore included those participants in all analyses.

The perceived validity of the IAT and participants' impression of their evaluator showed that, as intended, there were no reliable effects of experimental condition on participants' perceived validity of the test (overall M = 3.32, SD = 1.48; F's ≤ 2.71, p's ≥ .10) or their impression of their evaluator, which was quite positive overall ($M_{competent}$ = 5.17, SD = 1.14; M_{kind} = 5.70, SD = 0.87; M_{moral} = 5.24, SD = 0.97; all F's ≤ 3.87, p's ≥ .06).

IAT effect (*D* score).

An ANOVA with task domain, evaluator's minimal group and evaluator's religion as independent factors revealed a significant main effect of evaluator's religion, $F(1,215) = 11.68$, $p = .001$, $\eta_p^2 = .05$. Whereas participants whose evaluator was a non-Muslim woman showed significant bias against Muslim women (M = 0.16, SD = 0.45; $t[108] = 3.73$, $p < .001$), this bias was reduced to non-significance when participants were evaluated by a Muslim woman (M = -0.04, SD = 0.45, $t[113] = -0.90$, $p = .37$). Additionally, the interaction between task domain and evaluator's religion was marginally significant, $F(1,215) = 2.88$, $p = .09$, $\eta_p^2 = .01$. Analysis of simple main effects indicated that when evaluated by a Muslim woman there was no difference in IAT bias between the morality and competence condition (M = -0.03, SD = 0.50, M = -0.04, SD = 0.41 respectively; $F < 1$). However, when evaluated by a non-Muslim woman, participants for whom the moral implications of the test were emphasized showed a significantly weaker negative bias (M = 0.07, SD = 0.46) than participants for whom the implications of the test concerning their competence were emphasized (M = 0.27, SD = 0.42), $F(1,215) = 4.99$, $p = .03$, $\eta_p^2 = .02$. These results show that having a Muslim evaluator present is an impactful way of reducing non-Muslims' implicit anti-Muslim bias. However, even in the absence of an evaluator from the target group, a focus on morality rather than competence also reduces implicit bias significantly[10].

[10] A prior study (Van Nunspeet et al., under review) showed that emphasizing morality rather than competence reduced implicit bias in the presence of a (non-Muslim) evaluator belonging to a minimal ingroup, but not when this evaluator belonged to a minimal outgroup. Although this interaction effect was not significant in the current study ($F < 1$), the effect of task domain was indeed stronger when participants thought they were evaluated by a minimal ingroup member ($M_{morality}$ = 0.04, SD = 0.52; $M_{competence}$ = 0.27, SD

Inspection of reaction times.

To examine whether the effects of evaluator's religion and task domain on implicit bias were due to enhanced positive associations with the Muslim outgroup (reduced RTs on incongruent trials) or the inhibition of prepotent biased responses (increased RTs on congruent trials), we analyzed response latencies on correctly answered congruent and incongruent trials separately.

Congruent trials. The analysis of response latencies on correct *congruent* trials (reflecting the speed of making stereotype-congruent associations) revealed significant effects of our manipulations in line with the observed pattern of implicit bias reduction reported above. Parallel to the effect of evaluator's religion on the implicit bias score, evaluator's religion significant affected RTs on congruent trials, $F(1,215) = 7.09$, $p = .008$ $\eta_p^2 = .03$. Participants whose evaluator was a Muslim woman responded more slowly on congruent trials ($M = 503.97$, $SD = 24.24$) than participants whose evaluator was a non-Muslim woman ($M = 495.45$, $SD = 27.13$). Moreover, replicating previous work (Van Nunspeet et al., 2014), participants working under moral task instructions responded significantly more slowly on congruent trials ($M = 502.81$, $SD = 24.33$) than participants in the competence condition;($M = 496.88$, $SD = 27.30$), $F(1,215) = 3.92$, $p = .05$, $\eta_p^2 = .02$. Finally, participants responded marginally slower on congruent trials when their evaluator was a minimal ingroup member ($M = 502.89$, $SD = 26.64$) than when she was a minimal outgroup member ($M = 496.91$, $SD = 25.14$), $F(1,215) = 2.73$, $p = .10$, $\eta_p^2 = .01$.

Although there were no significant interaction effects; F's ≤ 1.84, $p \geq .18$, to enable a more direct comparison with the analyses for overall implicit bias, we analyzed RTs on congruent trials per evaluator's religion condition. Replicating the pattern for implicit bias, when participants were evaluated by a Muslim woman there were no significant effects of task domain or evaluator's minimal group on congruent response latencies (F's ≤ 2.44, p's $\geq .12$). However, when evaluated by a non-Muslim woman, participants responded significantly slower on congruent trials

---

$= 0.52$; $F[1,105] = 3.73$, $p = .06$, $\eta_p^2 = .03$), compared to a minimal outgroup member ($M_{morality} = 0.11$, $SD = 0.40$; $M_{competence} = 0.26$, $SD = 0.31$, $F[1,105] = 1.49$, $p = .23$).

in the morality condition (M = 500.61, SD = 23.66) than in the competence condition (M = 489.36, SD = 29.83), $F(1,105) = 4.67$, $p = .03$, $\eta_p^2 = .04$.

Incongruent trials. Analysis of response latencies on the correct incongruent trials (reflecting the stereotype-incongruent combinations of Muslims/positive and non-Muslim/negative) revealed no main effects of task domain, evaluator's religion or evaluator's group type, nor the interaction between evaluator's religion and task domain found for the overall D-score (all F's $\leq$ 1.04, $p \geq .31$). Thus, the experimental manipulations that resulted in a reduction of implicit bias did not cause participants to respond faster on incongruent trials[11].

Discussion

The results of Study 5.1 showed that participants reduced their anti-Muslim bias in case of presence of a Muslim evaluator or, in the absence of a Muslim evaluator, the emphasis on their morality instead of their competence. Moreover, this bias reduction was associated with the inhibition of stereotype conforming responses rather than with increased positive associations with the Muslim outgroup. Although these findings are consistent with previous research (Richeson & Ambady, 2003; Van Nunspeet et al., 2014), we wanted to test whether they are dependent upon the duration of the experiment: If positive associations have to be learned, they may only develop over a longer period of time.

We examined this possibility in Study 5.2, in which we increased the exposure to participants' evaluator while using the same cross-categorization dimensions as in Study 5.1. If participants share their minimal group membership with their Muslim evaluator, they may become to perceive their evaluator as a partial ingroup member when the duration of the interaction is increased (see also Crisp & Hewstone, 1999; Crisp, Hewstone, & Rubin, 2001, for effects of cross-categorization). Moreover, perceiving the evaluator as a partial ingroup member

[11] We also found an unexpected interaction between task domain and evaluator's group type, $F(1,215) = 4.02$, $p = .05$, $\eta_p^2 = .02$. Whereas there was no difference between the minimal group types of the evaluator in the morality condition ($M_{ingroup}$ = 495.28, SD = 24.98; $M_{outgroup}$ = 498.33, SD = 22.66, $F < 1$), participants in the competence condition responded faster on incongruent trials when the evaluator was a minimal outgroup (M = 490.10, SD = 21.74) instead of a minimal ingroup member (M = 499.61, SD = 22.78), $F(1,215) = 4.63$, $p = .03$, $\eta_p^2 = .02$.

may facilitate positive associations with the Muslim outgroup. In Study 5.2, we thus significantly increased the number of IAT trials to enable participants to develop new (positive) associations with Muslims during the task (resulting in reduced RTs on incongruent trials).

Study 5.2

Method

Participants.

Only female, non-Muslim, students (N = 102; M_{age} = 21.3 years, SD = 3.1) participated in the study for money or course credits. One participant was excluded from the analyses because she responded too late on more than 25% of the IAT trials, suggesting lack of attention to the experimental task.

Procedure.

The IAT and the procedure were similar to those described in Study 5.1. However, in Study 5.2, all participants received feedback from a Muslim evaluator. Thus, participants were randomly assigned to one of the four experimental conditions of the 2 (Task Domain: morality/competence) x 2 (Evaluator's Minimal Group: ingroup/outgroup) between-participants design. Moreover, the amount of trials in the two test blocks of the IAT was increased: From 70 trials per block in the previous study to 120 trials per block in Study 5.2.

Results

Checks.

Ninety-eight percent (N = 49) of participants in the morality condition and 96% (N = 49) of participants in the competence condition correctly reported the task domain. Moreover, 92% (N = 47) of participants whose evaluator was an ingroup member and 98% (N = 49) of participants whose evaluator was an outgroup member reported their evaluators' minimal group correctly. Because exclusion of the participants who answered one of the checks incorrectly (N = 6) did not alter the pattern of means, we included those participants in all analyses.

As intended, participants in all four conditions indicated that the test was able to assess what kind of person they are to a similar degree; overall M = 3.44, SD = 1.57; F's ≤ 1.23, p's ≥ .27. Moreover, there were no effects of our task domain or evaluator's minimal group manipulation on participants' impression of their

evaluator, which was quite positive overall ($M_{competent}$ = 5.36, SD = 1.09; M_{kind} = 5.85, SD = 0.84; M_{moral} = 5.54, SD = 0.98; all F's ≤ 1.68, p's ≥ .20).

IAT effect (*D* score).

Consistent with Study 5.1, now that all participants were evaluated by a Muslim woman, on average they did not show implicit bias against Muslim women, M = -.02, SD = .32, $t(100)$ = -.53, p = .60. Additionally, an ANOVA with task domain and evaluator's minimal group type as independent factors revealed a main effect of evaluator's minimal group type: Participants whose Muslim evaluator was presented as a minimal ingroup member showed significantly less bias against Muslim women (M = -0.08, SD = 0.27) compared to participants who thought they were evaluated by an outgroup member (M = 0.05, SD = 0.35), $F(1,97)$ = 5.02, p = .03, η_p^2 = .05. The effect of task domain was marginally significant, $F(1,97)$ = 2.89, p = .09, η_p^2 = .03: In line with the previous findings the means show that implicit bias was reduced under moral task instructions (M = -0.07, SD = 0.33) compared to competence instructions (M = 0.03, SD = 0.29).

We proceeded by examining whether RTs on correct congruent and incongruent trials differed across experimental conditions. Interestingly, the general tendency to slow down on congruent trials indicating the inclination to inhibit prejudice conforming responses did not depend on the evaluator being an in- or an outgroup member or on task domain (F's ≤ 2.66, p's ≥ .11). Additionally, and as expected, we found evidence in line with our reasoning that increasing the number of trials in which participants are exposed to a Muslim evaluator who is presented as an ingroup member can facilitate the ability to associate positive stimuli with Muslim targets. That is, participants responded faster on incongruent trials when the Muslim evaluator was presented as a minimal ingroup member (M = 478.87, SD = 23.90) than when she was an outgroup member (M = 493.26, SD = 23.08), $F(1,97)$ = 9.47, p = .003, η_p^2 = .09[12]. This suggests that the decrease in implicit bias

[12] To directly test the effect of the increase in trials, we combined the data of Study 5.2 (N = 101) with the data of participants who were evaluated by a Muslim evaluator in Study 5.1 (N = 114). Results of an ANOVA with RTs on incongruent trials as dependent variable and amount of trials, task domain and evaluator's minimal group type as independent factors showed a main effect of amount of trials: Participants responded significantly faster on incongruent trials when the amount was increased

observed when the Muslim evaluator was a minimal ingroup member reflects that the ability to associate Muslim individuals with positive stimuli is facilitated under these conditions.

General Discussion

In the current research we directly compared the effects of three different interventions on people's implicit evaluative bias against Muslims: (1) People's personal motives to appear moral; (2) their intergroup motivation to perform well towards a Muslim evaluator, and (3) their intragroup-based motives to perform well in front of self-relevant others (categorized on a second, minimal, group dimension). We tested these effects by introducing a Muslim/non-Muslim IAT as a measure of participants' moral values or of their competence. Moreover, participants performance was evaluated by either a non-Muslim or Muslim individual who was presented as a minimal in- or outgroup member. Results of Study 5.1 revealed the significant effect of target presence: In line with previous research (Lowery et al., 2001), participants showed no sign of anti-Muslim bias when they their evaluator was Muslim. Moreover, the significant reduction in bias was associated with the inhibition of prejudice: Instead of decreased response times on incongruent trials (indicating rapid associations between Muslims and positive attributes and non-Muslims and negative attributes), participants slowed down their responses on congruent trials, suggesting that they aimed to inhibit their prepotent responses to rapidly associate Muslims with negativity and non-Muslims with positivity.

In case participants' evaluator was not Muslim, we did find the same pattern of inhibition of prejudice-conforming responses when the moral implications of the test were emphasized: When participants were told that their test score could be perceived as an indication of their moral values concerning egalitarianism, this helped them to show a smaller bias against Muslims than when they were told that

---

$M_{120trials}$ = 485.99, SD = 24.47, $M_{70trials}$ = 496.90, SD = 22.66, F(1,207) = 11.82, p = .001, η_p^2 = .05. Moreover, there was a significant interaction effect between amount of trials and evaluator's minimal group type (F[1,207] = 7.50, p = .007, η_p^2 = .04), indicating that participants only responded faster on incongruent trials while they were evaluated by a minimal ingroup member in case of the increased amount of trials; $M_{120trials}$ = 478.86, SD = 23.90, $M_{70trials}$ = 498.75, SD = 22.64, F(1,207) = 18.82, p < .001, η_p^2 = .08.

their test could reveal their competence. Emphasizing one' morality thus seems to be an effective way to facilitate bias reduction and may be an alternative intervention when intergroup contact is not feasible.

Furthermore, in line with previous research (e.g., Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000), results of Study 5.2 revealed that new (positive) associations can also be induced. First of all by increasing the amount of exposure to a Muslim evaluator and thus by emphasizing that one's intergroup behavior is evaluated by an outgroup member. And second, by introducing cross-categorization and focusing people on what they have in common with someone who they perceive as an outgroup member on another social dimension: Presenting a Muslim (outgroup) evaluator as a minimal ingroup member helped participants to developed positive associations with the Muslim outgroup. Importantly, our results extend prior research which revealed that shared (minimal) group membership(s) can override people's explicit evaluative bias against outgroup members (e.g., Crisp et al., 2001; Urada, Stenstrom, & Miller, 2007), by showing similar findings for people's implicit bias.

Our findings indicate that there are different ways in which implicit prejudice can be reduced. The presence of a member of the target outgroup may have the greatest impact on the control of prejudiced responses and can even activate new (positive) associations with the outgroup. However, we should not overestimate this effect in everyday interactions: Social groups that are the focus of prejudice research are generally minority groups in society that are often segregated from the majority in education, housing, and work, preventing extensive intergroup interactions. The current research thus offers a contribution to insights on prejudice reduction by demonstrating again the potential impact of emphasizing one's morality and the presence of others who share the same ingroup norms, even when no outgroup member is present (see also Van Nunspeet et al., 2014).

We note that specific circumstances were in place in the current research as it remains unclear which aspect of our manipulations concerning the Muslim evaluator caused the effect of faster positive associations with Muslim women. Our participants received feedback on every trial and since they made few errors, they received almost continuous positive feedback. They thus repeatedly saw a smiling,

approving Muslim woman who was presented as someone like them (an ingroup member). It is less likely that similar effects will be obtained when participants were provided with as much or more negative rather than positive feedback.

Nevertheless, we have shown that evaluative bias against Muslims can be reduced by several means. Presence of a Muslim evaluator causes people to inhibit their prejudiced responses and, provided there is enough exposure, presenting her as a self-relevant other may strengthen positive associations. Moreover, besides this form of intergroup contact, prejudice control can also be instigated by emphasizing people's moral values.

Acknowledgements

We thank Johannes Parzonka, Eva Schildkamp, and Titia van Malestein for help with the data collection; Bianca van Nunspeet for her contribution to the experiment; and Kees Verduin for his help with processing the data.

Part III

The need for confirmation of one's own morality

Chapter 6

Affective and attentional responses to positive and negative feedback about one's own moral behavior

Collaborators on the research described in this chapter are:
Naomi Ellemers, Eveline Crone, Belle Derks, and David Amodio. Their contribution can be specified as follows: Design of the studies: FvN, NE, EC, BD, DA. Performing the experiments: FvN. Analysis of skin conductance data: FvN. Analysis of fMRI data: FvN, EC. Writing of the paper: FvN, NE, EC, BD. *Manuscript in preparation.*

A general principle in psychology is that bad is stronger than good (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). Although applicable to many types of judgments and situations, this is also an essential mechanisim in judging someone else's moral integrity. This was established by Skowronski and Carlston (1987), who examined positive and negative extremity biases for morality and competence judgments during impression formation. Their findings revealed that negative rather than positive behaviors are perceived as more diagnostic for someone's 'true character' when these refer to the moral domain. In contrast, however, positive rather than negative behaviors are perceived as more diagnostic for someone's personality when these behaviors relate to their competence. In other words, we assume that everyone can act in a moral way, for instance when criminals pretend to be upright citizens - so this is non-diagnostic. However, only *im*moral people should do immoral things. Conversely, we tend to think that everyone can do something *in*competent once in a while – even a professor can be confused or forgetful - but only competent people should be able to behave competently.

This negative extremity bias concerning morality (e.g., Lupfer, Weeks, & Dupuis, 2000) and the differential diagnosticity of moral and competent behaviors (e.g., Martijn, Spears, Van der Pligt, & Jakobs, 1992) have been observed in empirical research. However, prior studies have focused on impression *formation* about others – examining this from a perceiver's perspective. Thus far, it has remained unclear whether a similar asymmetry in the value attached to moral vs. competent behaviors is also evident in impression *management* – in the concerns people have about the image of the self in the eyes of others (from an actor's perspective). To the extent that positive and negative extremity biases for morality and competence are also associated with impression management about the self, people should be strongly preoccupied with avoiding to display any behavior that might indicate their immorality, and focus on providing confirmation of their competence. Because it is not always possible to act in line with one's ambitions and ideals, people are likely to be confronted from time to time with others who provide negative evaluations of their moral or competent behavior. We argue that the asymmetrical implications of person information concerning morality vs. competence should therefore be visible in the affective states people experience.

That is, they should suffer increased negative states when being confronted with negative information concerning their own morality (as compared to their competence). Conversely, they should experience increased positive states when they receive information about their own competence (as compared to their morality). Previous research concerning people's self-perceptions and impression management has revealed evidence offering partial support for this reasoning, as it has established that people tend to attach greater importance to moral information about the individual or group self then to competence information. That is, overall people indicate they perceive moral traits as more important characteristics of their personal and social identity than traits referring to their competence (and sociability; Leach, Ellemers, & Barreto, 2007). They indicate being motivated to display behavior that is seen as moral as a way to secure inclusion in a group (Ellemers, Pagliaro, Barreto & Leach, 2008) and to earn respect from fellow ingroup members (Pagliaro, Ellemers, & Barreto, 2011). Moreover, this motivation to display moral behavior is also evident at a less explicit level as people tend to inhibit their social bias against Muslims (i.e., display a moral task performance) when the test used to assess this was said to be indicative of their morality instead of their competence (Van Nunspeet, Ellemers, Derks, & Nieuwenhuis, 2014).

Prior research thus underlines the importance of morality over competence in impression management about the self. This is the case when people have to explicitly state their preference or when they are assigned to a task condition that emphasizes either moral or competence implications of task performance. As yet, it still needs to be examined whether the greater value attached to moral information about the self relates to the desire to avoid appearing immoral, or stems from the ambition to demonstrate one's ability to behave morally. The aim of the present research was to directly compare the impact of these different types of information related to the self, as a way to establish whether people differentially welcome information that might confirm their morality or competence in a positive way, or are disturbed by negative information depending on whether it threatens to reveal their lack of morality or competence.

One way of examining the impact of different types of information supposedly relevant to the self, is to ask participants to report how they feel after

receiving this information. Such a method relies on the introspective capabilities of participants and may be affected by people's explicit preferences for a particular type of information over the other, as well as their willingness to reveal these to the experimenter. Thus, such self-report measures do not necessarily provide a reliable picture of their internal states. Psychophysiological measures seem to offer a solution for these difficulties associated with self-report measures. For example, electrodermal activity, often measured as skin conductance, is an automatic response from the sympathetic nervous system caused by arousing stimuli (for an overview see Dawson, Schell, & Filion, 2000). Indices of skin conductance can thus not easily be adapted by the participant for self-presentational reasons, and can be measured online (i.e., to monitor changed states *while* participants receive relevant information, instead of relying on retrospective reports). Combining self-reports with skin conductance data can thus elucidate how people respond to information about their own behavior and compare this to what they report when thinking back about the information.

In addition, previous neuroscientific research has been able to disentangle different cognitive processes associated with processing self-relevant information (most often using functional magnetic resonance imaging; fMRI). That is, processes associated with the *detection* of self-relevant information seems to be associated with different parts of the brain (i.e., the ventral medial prefrontal cortex, vMPFC) than the *evaluation* of self-relevant information (i.e., the dorsal medial prefrontal cortex, dMPFC; for reviews, see for example Northoff & Bermpohl, 2004; Van der Meer, Costafreda, Aleman, & David, 2010). Comparing fMRI responses observed in these two areas allows us to establish the extent to which people detect information as being self-relevant, and separate this from their tendency to relate this to their actual self-views. In the current research, we thus combined these different indicators of the way participants process self-relevant information: We measured participants' self-reported affective reactions *after* having received either positive or negative feedback about their scores on a measure of their morality and competence. In addition, we measured their skin conductance to assess physiological arousal (Study 6.1) and used fMRI to examine mental processing (Study 6.2) *while* receiving morality and competence feedback.

Mental Processing of Self-relevant Information

Previous neuroscientific research has examined the neural networks involved in processing information relevant for the self. Prior research has addressed the brain regions involved in the assessment of self-relevant information (i.e., processing information that people perceive as related to the self; Northoff & Panksepp, 2008; Schmitz & Johnson, 2007), and reported networks including both subcortical and cortical regions (e.g., caudate nucleus, amygdala, Insula, and anterior singulate cortex [ACC]; Schmitz & Johnson, 2007). Moreover, there is high consensus on the role of the medial prefrontal cortex (MPFC) in processing such self-relevant information (e.g., Abraham, 2013; Ochsner et al., 2005; Northoff & Bermpohl, 2004; Schmitz & Johnson, 2007). In fact, Moran, Macrae, Heatherton, Wyland, and Kelley (2006) showed that MPFC activation during self-referencing was affected by self-relevance. That is, activation in the MPFC was greater when participants judged personality characteristics (i.e., traits words such as "honest") as high self-relevant as compared to low self-relevant. In line with these findings, and given that we expect that information concerning morality is more self-relevant than information concerning competence, we will examine whether receiving feedback about one's morality is associated with greater activation in the MPFC than receiving feedback about one's competence. appraisal

Although activation in the MPFC is found in many studies concerning self-relevance in general, subregions within the MPFC seem to be associated with more specific processes. For example, in their review, Amodio and Frith (2006) discuss that whereas the posterior rostral region of the MFC is activated during action-monitoring tasks, the anterior rostral MFC is activated during tasks involving self-knowledge, person perception and mentalizing. Moreover, Van der Meer et al. (2010) made a distinction between the ventral and dorsal part of the MPFC and argued that the vMPFC is associated with detecting and labeling self-relevant information, and the dMPFC with evaluation and decision-making processes in self-referential thinking. In the current research, in which participants are only asked to passively view their scores on a measure indicative of their moral and competence behavior, we hypothesize that information concerning morality will be

perceived as more self-relevant than information concerning competence which could thus be associated with activation in the ventral MPFC.

Current Research

The current research aims to investigate whether the differential diagnosticity of morality and competence that is found in impression formation of others is also evident when people are informed about their *own* morality and competence. Based on social psychology research, which has shown that people perceive moral traits as more significant for their social and personal identity than traits concerning competence (e.g., Leach et al., 2007; Ellemers et al., 2008), we predict that receiving information concerning one's own morality (as compared to one's competence) is associated with increased self-reported emotional responses, arousal (assessed by a measure of skin conductance) and greater activation in the MPFC. In addition, impression formation research (e.g., Skowronski & Carlston, 1987) has revealed that negative, rather than positive, information is perceived as a better indication of someone's moral integrity. Conversely, positive rather than negative, information tends to be perceived as a better indication of someone's competence. Drawing on these findings relating to impression formation of others, we predict parallel effects when people receive evaluative information about the self. This is why we anticipate the valence of self-related information to interact with the dimension (competence vs. morality) to which this information pertains.

Study 6.1

Method

Participants.

Thirty three students (six males, M_{age} = 18.9, SD = 1.45) from Leiden University participated in the study in return for course credits or money. Five participants were not included in the SCR data analyses because of technical failures in the equipment or software; three other participants were excluded from the SCR data analyses because the signal was extremely noisy, and one other participant was excluded from the SCR data analyses since we could not measure a skin conductance signal. Participants were randomly assigned to one of two conditions: They either received positive or negative feedback (i.e., measured between participants) concerning their morality and competence (measured as a

within-participants factor). To enhance the credibility of the feedback provided, in both experimental conditions the valenced feedback was interspersed with evaluatively neutral feedback.

Procedure.

The feedback participants received was said to be based upon their performance on an Implicit Association Test (IAT; Greenwald et al., 1998) which participants completed in the first part of the experiment. The (non-) Muslim IAT in the current study has previously been used to examine whether people adjust their performance when the test is presented as indicative of their morality (i.e., by informing participants that the test can assess their moral values concerning egalitarianism and discrimination) or of their competence (i.e., by informing participants that the test can assess their ability to process information and learn new tasks; Van Nunspeet et al., 2014). Moreover, since this previous research has shown that participants indeed perceive the test as a credible measure of both properties, we implemented the IAT in the current research as a task on which we could present participants with feedback about their moral values as well as their competence in displaying accurate responses. Importantly, in the current study, participants were informed about these test implications *after* they had finished the IAT, right before they received their feedback to keep task motivation and effort constant across experimental conditions.

The IAT included pictures of female faces with and without a headscarf that had to be associated with positive and negative images (International Affective Picture System; Lang et al., 2005). Congruent IAT trials were trials on which participants were asked to press one response key when viewing both female faces with a headscarf and negative pictures and another key when viewing female faces without a headscarf and positive pictures. Incongruent trials were trials on which the same response key had to be pressed for pictures of female faces with a headscarf and positive pictures and another key when viewing female faces without a headscarf and negative pictures. In order to present participants with several instances of feedback (i.e., necessary for reliable skin conductance data), they performed 20 test blocks of the IAT; each test block consisted of eight trials.

After the IAT participants were informed about the implications of the test. That is, they were led to believe that the test is able to assess both their level of competence (tested as their ability to quickly process new information and to learn new tasks), as well as their level of morality (i.e., their moral values concerning egalitarianism and discrimination). Moreover, participants read that their scores on these two test domains would be provided relative to the scores of other university students and could thus give an indication whether they had scored above average (positive feedback, indicating relatively high moral values or competence), below average (negative feedback, indicating relatively low moral values or lack of competence), or whether their scores were average for the student population (neutral feedback). Neutral feedback was included to enhance credibility of the cover story, and as a control - to be able to check whether above or below average scores affected participants more than average (evaluative neutral) scores. The valence of the feedback was manipulated between-participants - since we did not think it would be credible to provide participants with both above and below average scores on a single measure.

Scores were preprogrammed and represented by colored bars in a normal distribution in which the right hand side displayed above average scores related to morality (or competence) and the left hand side below average scores related to immorality (or incompetence). The participant's score was indicated by a red (negative), green (positive) or yellow (neutral) bar in the normal distribution and the text "your score" right above it (see Figure 6.1).

Participants either received positive (and neutral) or negative (and neutral) feedback. Each round of feedback was provided in two blocks in which one block concerned feedback related to one's morality and the other block feedback related to one's competence. Before each block, participants read the information concerning the nature of the task domain under examination (competence or morality). The order of the feedback blocks was counterbalanced between participants. Each block consisted of ten rounds of valenced (positive or negative) feedback interspersed with ten rounds of neutral feedback. Every feedback round consisted of a screen stating that participants' next test score (concerning their morality or competence) was being computed (9 - 11 sec.), followed by a screen

providing the presentation of the feedback (3 sec.). After viewing their test score for three seconds, participants could press a key to go to the next round of feedback (see Figure 6.1).

Skin conductance was assessed during the IAT as well as the feedback phase to enable participants to get used to the equipment that was attached and to avoid drawing particular attention to a particular part of the experiment as being of special interest. After completing the IAT and before the feedback was provided, the waiting time was used to derive a baseline measure for skin conductance. After having received all the feedback, participants were asked to complete some self-report questionnaires (see details below). The experiment lasted approximately thirty minutes in total, after which participants were properly debriefed about the bogus feedback and the actual goal of the study. They were then thanked and received their incentive.

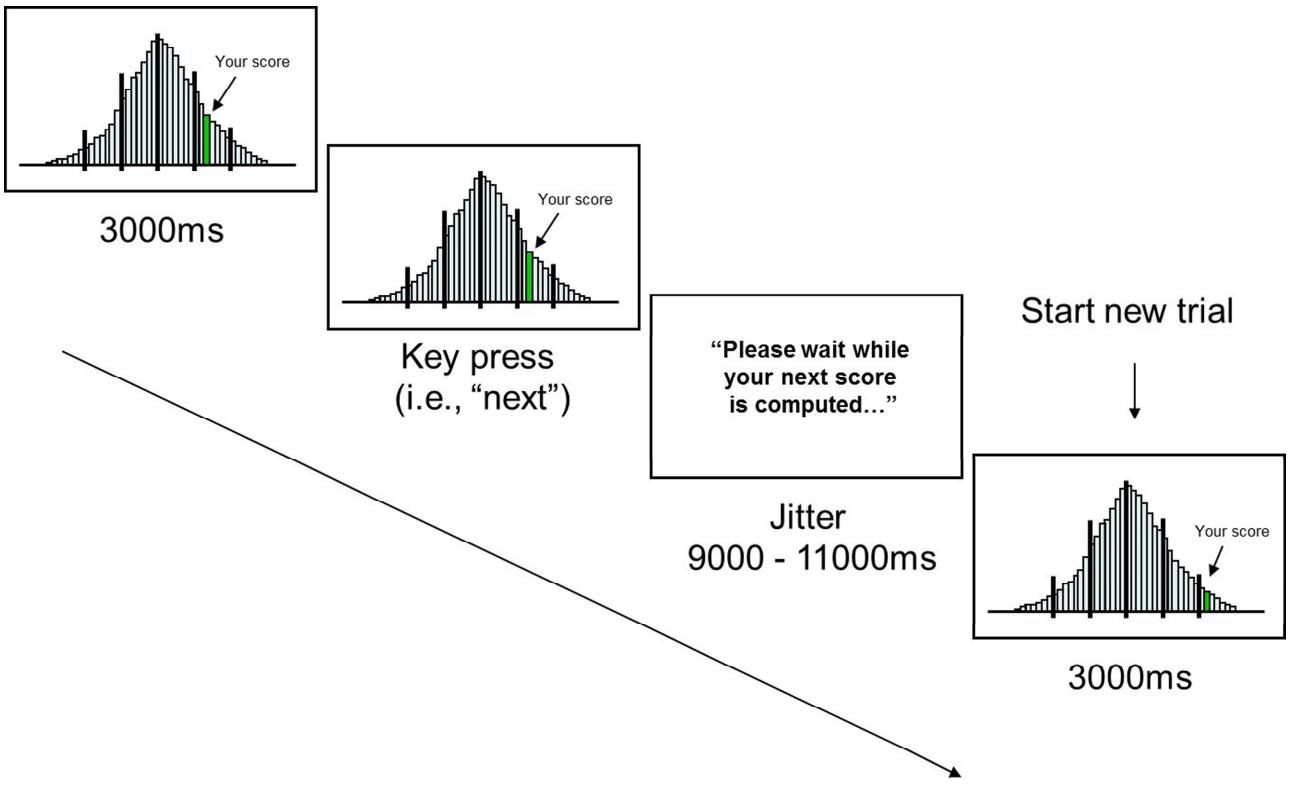

Figure 6.1. Example trial of feedback presented in Study 6.1.

Skin conductance acquisition and processing.

Skin conductance was measured using two pregelled disposable Ag-AgCl electrodes attached to the medial phalanx surfaces of the middle and index fingers of the non-dominant hand. The transponder unit relayed skin conductance data to a host computer running AcqKnowledge software, which logged every feedback stimulus-onset on the skin conductance signal. The data were filtered online with a low pass filter of 2 Hz and offline with a low pass filter of 0.33 Hz. The data was processed in two ways: We measured whether the feedback resulted in an elevated skin conductance level (SCL) compared to baseline, and we determined whether each feedback trial resulted in elevated skin conductance responses (SCRs). For the first measure, we computed difference scores between the average SCL in a 0-6 seconds time window after stimulus-onset in comparison to the average SCL in the final 30 seconds of the baseline measure, separately for each type of feedback (i.e., neutral and valence feedback concerning morality and competence). For the second measure we detected SCRs with a minimum amplitude change of 0.01 μS after stimulus-onset, and measured the number of SCRs in a time window between 1 and 6 seconds after each stimulus-onset. When there was no SCR associated with the feedback-stimulus, "0" was recorded. The mean number of SCR's was then calculated separately for feedback indicating scores on morality and competence and separately for neutral and valenced feedback. It should be noted that since many participants failed to generate SCR's related to the feedback, the mean number of fluctuations was below 1.0 (which is in line with previous research; e.g., Lawrence et al., 2006).

Self-reports.

Checks. We first checked whether participants had experienced the task in a similar way and were equally uncertain about their performance, regardless of whether they had received positive or negative scores. For this purpose, after having received all of their feedback, we asked participants to answer two questions about their experience while performing the IAT (i.e., "I was insecure about my performance on the test" and "During the test, I had the feeling I was able to perform very well" [recoded], $r = .44$, $p = .011$). They could indicate their answers on a 7-point Likert scale (1 = completely agree – 7 = completely disagree).

Self-reported negative emotional response. We then asked participants to reconsider how they felt while receiving the feedback, and to indicate the emotional response this raised. Items asked participants to indicate their general feelings (i.e., "Seeing my scores gave me a bad feeling"; "My scores gave me the idea that I don't have good qualities"; "Seeing my scores gave me a good feeling"; "My scores made me feel good about myself") as well as a number of specific emotions (i.e., "When I received feedback concerning the morality/competence domain of the test, I felt: discouraged / nervous / guilty / ashamed / threatened / frustrated / happy / relaxed / motivated / proud / enthusiastic / challenged"). All answers were assessed using 7-point Likert scales (1 = completely agree – 7 = completely disagree). All these questions were asked twice: Once to indicate emotional responses to morality feedback and once to convey emotional responses to competence feedback. Items concerning positive feelings and emotions were recoded so that higher scores always indicated a more negative emotional response. We then combined the items concerning general feelings and specific emotions for each type of feedback, resulting in two overall indicators. One combined score indicated the degree to which participants reported a negative emotional response when viewing their scores on morality (α = .94) the other indicated negative emotional responses when viewing their competence scores (α = .91).

Results

Checks.

To check whether participants were equally uncertain about their task scores so that the feedback they received seemed credible regardless of experimental condition, we asked participants to indicate their thoughts about their performance during the IAT. Results of a one-sample T-test with the mean of the scale (4) as the test value showed that, overall, participants reported to be quite insecure about their performance (M = 4.77, SD = 1.29; t[32] = 3.45, p = .002). There were no differences between experimental conditions, suggesting that below or above average test scores would seem equally plausible.

Skin conductance data.

Skin conductance level (SCL). To test whether the feedback presented during the experiment affected participants' arousal levels (irrespective of valence

or task domain), we first tested the difference between the average SCL following the feedback (i.e., 0-6 seconds after stimulus-onset, across all types of feedback) and the average SCL during the final 30 seconds of the baseline. Results of a paired sample T-test revealed that, as intended, the feedback significantly increased SCL as compared to baseline, $M_{difference} = 0.64$, $SD = 1.38$, $t[23] = 2.26$, $p = .03$.

To examine any differences in SCL between the types of feedback, we conducted a repeated measures ANOVA on the difference scores of SCL (0-6 seconds after stimulus-onset minus baseline) with the type of feedback (valenced/neutral) and task domain (morality/ competence scores) as repeated measures, and the context in which feedback was provided (positive/ negative feedback condition) and order (morality/competence block first) as between-groups factors. Results revealed a significant main effect of feedback type; $F(1, 20) = 11.45$, $p = .003$, $\eta^2_p = .36$, indicating that SCL was greater after valenced ($M = 0.68$, $S.E. = 0.31$) compared to neutral feedback ($M = 0.56$, $S.E. = 0.30$). This main effect was however qualified by a significant feedback type*order interaction effect; $F(1, 20) = 7.46$, $p = .01$, $\eta^2_p = .27$, revealing that the difference between valenced and neutral feedback was only significant when the scores concerning morality were presented first; $F(1, 20) = 16.33$, $p = .001$, $\eta^2_p = .45$. The other simple main effects were not significant; all F's < 1. There were no interaction effects with task domain, indicating that there were no differences in average SCL between positive/negative or neutral feedback related to morality and competence.

Skin conductance responses (SCRs). To examine whether the different types of feedback affected skin conductance directly after stimulus-onset, we also analyzed SCRs. We assessed differences in SCRs during the feedback round with a repeated measures ANOVA with the type of feedback (valenced/neutral) and task domain (morality/ competence scores) as repeated measures, and the context in which feedback was provided (positive/ negative feedback condition) and task domain (morality/competence scores) as repeated measures, and valence (positive vs. negative feedback) and order (morality vs. competence block first) as between-groups factors. Results revealed no difference in SCR´s between valenced and neutral feedback; $F(1,20) = 2.32$, $p = .14$. However, we found evidence in support of our central prediction, indicating that feedback relating to morality had a greater

impact than feedback referring to competence: We observed a marginally significant main effect of task domain; $F(1,20) = 3.90$, $p = .06$, $\eta^2_p = .16$, indicating that there were more SCRs when participants were confronted with their morality ($M = 0.36$, $S.E. = .03$) than competence scores ($M = 0.29$, $S.E. = 0.04$). This effect was qualified by a significant interaction effect between task domain and order; $F(1,20) = 5.19$, $p = .03$, $\eta^2_p = .21$, indicating that a significant difference in SCRs between morality and competence feedback only emerged when the morality scores were presented first (i.e., increased SCR's in the morality [$M = 0.40$, $S.E. = .05$] compared to the competence block [$M = 0.21$, $S.E. = .08$], $F[1,20] = 7.90$, $p = .01$, $\eta^2_p = .28$). When competence scores were presented first there was no difference in responses to the different task domains ([$M_{morality} = 0.31$, $S.E. = .04$; $M_{competence} = 0.32$, $S.E. = .07$]; $F < 1$). Additionally, we observed a trend towards a three-way interaction between task domain, order and valence; $F(1,20) = 2.99$, $p = .10$, $\eta^2_p = .13$. Examination of the repeated measures ANOVA separately for the positive and negative feedback conditions revealed that the task domain x order interaction effect could only be traced to the negative feedback condition; $F(1,11) = 7.36$, $p = .02$, $\eta^2_p = .40$, but not the positive feedback condition ($F < 1$; see Figure 6.2).

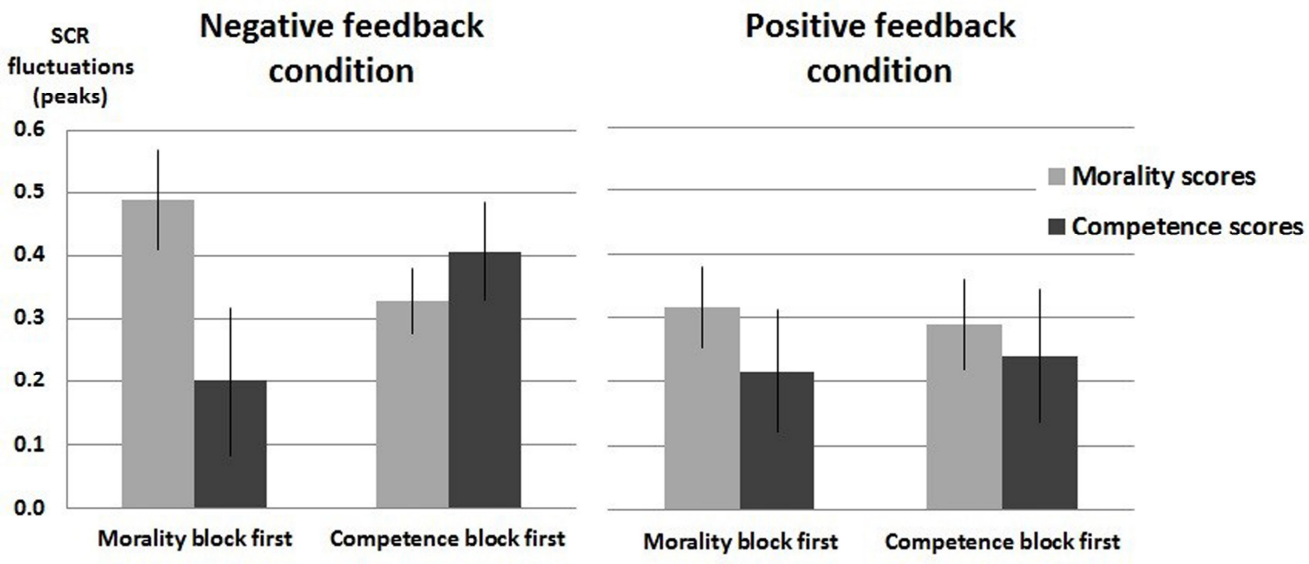

Figure 6.2. Average skin conductance responses (SCRs) in each condition. Whereas there were no differences in SCRs for positive feedback (right), negative feedback concerning morality was associated with increased physiological arousal –in case morality scores were presented first (left).

Self-reported negative emotional response.

After participants had received all of their feedback, we asked them to think back about the moments they received feedback about their morality and competence and to recall and report their emotional response. A repeated measures ANOVA with task domain (morality/competence) as the repeated measure and valence (positive/negative feedback)[13] as between-participants factor, revealed evidence in support of our reasoning. We observed a significant interaction effect between task domain and valence; $F(1,31) = 4.00$, $p = .05$, $\eta^2_p = .11$. The relevant means and analysis of simple main effects confirmed that the difference between positive and negative feedback conditions in self-reported emotional response was more pronounced when participants received feedback regarding their morality; $M_{\text{difference}} = 1.82$, $S.E. = 0.24$; $F(1,31) = 60.02$, $p < .001$, $\eta^2_p = .66$, rather than their competence; $M_{\text{difference}} = 1.44$, $S.E. = 0.22$; $F(1,31) = 42.37$, $p < .001$, $\eta^2_p = .58$. Specifically, when participants had received negative feedback they reported a more negative emotional response when the feedback was related to their morality ($M = 3.45$, $S.E. = 0.17$) rather than their competence ($M = 4.07$, $S.E. = 0.16$); $F(1,31) = 6.95$, $p = .01$, $\eta^2_p = .18$. There was no difference between responses to positive feedback depending on whether this pertained to the morality or the competence domain ($F < 1$).

Taken together, the findings of Study 6.1 offer evidence in line with our reasoning, as they suggest that receiving information related to one's morality has more impact on participants' responses than feedback related to their competence, in particular when people are confronted with negative feedback. To examine whether feedback concerning one's morality (as compared to competence) is also processed differently in the brain, we conducted an fMRI study in which we examined the neural network involved in processing self-relevant information.

[13] Note that we did not include a factor distinguishing between valenced and neutral feedback in this analysis, because we asked participants how they felt about their feedback overall, which was predominantly negative (negative and neutral) in the negative feedback condition, and predominantly positive (positive and neutral) in the positive feedback condition.

Study 6.2

Method

Participants.

Forty right-handed students (12 males, M_{age} = 21.7 years, *SD* = 3.1) from Leiden University participated in the study in return for course credits or money. None of the participants reported a history of psychiatric or neurological disorders, and current use of any medications. One participant was excluded from the analysis of the behavioral data because she failed to detect the color change of the fixation cross (whereas all other features of the stimuli were clear). Three other participants could not be included in the fMRI analyses because of technical problems. Participants were randomly assigned to the positive or negative feedback condition. All procedures were approved by the medical ethical committee of the Leiden University Medical Center (LUMC) and all participants gave informed consent for the study.

Procedure.

Before the scanning session, participants performed the (non-)Muslim IAT without receiving any information about the implications of the test, similar to Study 6.1. During the scanning session, participants were first informed that the test was able to assess both their level of competence, as well as their level of morality. In contrast to Study 6.1, participants thus read about *both* types of implications before they received any of the feedback stimuli. Participants were presented with the same feedback stimuli as used in Study 6.1.

Participants were informed about both types of test implications at once because the current study used an event-related block design: Feedback was provided in one run in which 6 blocks of feedback concerning morality were alternated with 6 blocks of feedback concerning competence. Each block consisted of 5 feedback trials of which two or three trials provided valenced feedback (positive or negative, depending on experimental condition) and two or three trials provided neutral feedback. The reason for presenting the competence and morality trials in mini blocks was to ensure direct repetition of each task domain, in order for the feedback to have impact on participants (which was similar to the block

design used in Study 6.1). In total there were 15 trials per feedback type (morality-valence/morality-neutral/competence-valence/competence-neutral)[14].

Each feedback round consisted of a screen stating that participants' next test score concerning their morality or competence was being computed (2 sec.), a fixation cross (jittered duration, 4-8 sec.), and the feedback stimulus (3 sec., see Figure 6.3). To ensure that participants were attentive, they were asked to press a key (with their right index finger) whenever the fixation cross changed color, which happened randomly after 1 to5 seconds.

As part of a larger study, the scanning session lasted approximately one hour. After the scanning session had ended, participants were asked to fill out some questionnaires. The complete study lasted approximately 2 hours, after which participants were properly debriefed, thanked and given their incentive.

fMRI data acquisition and processing.

Scanning was performed at the Leiden University Medical Centre (LUMC) with a standard whole-head coil on a 3.0 Tesla Philips Achieva scanner. Using E-prime 2.0 software, the task instructions and feedback was projected onto a screen at the back of the scanner bore, which participants could view via a window attached to the top of head coil. Participants could respond by pressing a button (using their right index finger) on a box attached to their right leg. The feedback was provided in one run, lasting approximately 15 minutes. Functional data were obtained using T2*-weighted echo-planar imaging ([EPI], repetition time (TR) = 2200 ms, echo time (TE) = 30 ms, slice matrix = 80 x 80, slice thickness = 2.75 mm, slice gap = 0.28 mm, field of view [FOV] = 220 mm). A high-resolution T2-weighted anatomical scan (same slice prescription as EPI) was collected at the end of the scanning as well as a high resolution 3D T1-weighted anatomical image (TR = 9.751 ms, TE = 4.59 ms, flip angle = 8°, 140 slices, 0.875 mm x 0.875 mm x 1.2 mm, and FOV = 224.000 x 168.000 x 177.333).

[14] The order of the blocks of feedback was not counterbalanced between participants (i.e., the first five feedback trials always concerned participants' morality and the following five participants' competence), which could have affected the results. We therefore also analyzed the data without the first ten trials to control for the possible high impact of these initial scores. Results of this analysis were similar to the ones described in the current results section.

Data were preprocessed and analyzed using SPM8 software (Welcome Department of Cognitive Neurology, London) implemented in MATLAB (Mathworks, Sherborn, MA). The functional time series were realigned to compensate for small head movements.

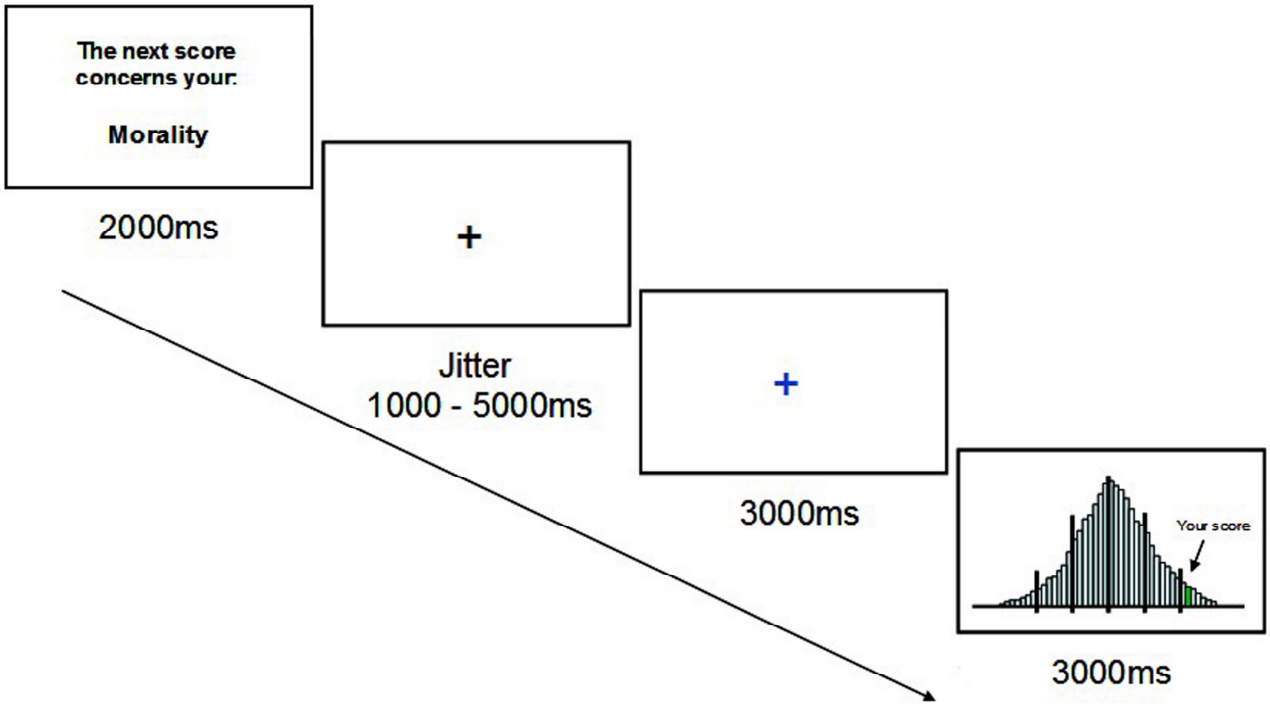

Figure 6.3. Example trial of feedback presented in Study 6.2.

Translational movement parameters never exceeded 1 voxel (< 3 mm) in any direction for any subject or scan. Functional volumes were spatially normalized to EPI templates. The normalization algorithm used a 12 parameter affine transformation together with a nonlinear transformation involving cosine basis functions and resampled the volumes to 3 mm cubic voxels. Functional volumes were spatially smoothed using an 8 mm full-width half-maximum Gaussian kernel. Templates were based on the MNI305 stereotaxic space (Cocosco, Kollokian, Kwan, Pike, & Evans, 1997), and the Montreal Neurological Institute (MNI) atlas was used to refer to the coordinates.

To analyze the data, a canonical hemodynamic response function was convolved at the onset of the feedback stimulus and modeled as a zero-duration event. We distinguished between four conditions *within* participants: Valence versus neutral feedback and feedback related to morality or competence. Whether the valence was positive or negative was a *between*-participants manipulation. These conditions resulted in four 2 X 2 full factorial designs. Two designs were used to examine the effects of valenced and neutral feedback for the positive and negative feedback conditions separately, resulting in two 2 (Feedback: Valence/Neutral) X 2 (Task Domain: morality/competence) ANOVAs which were run separately for the positive feedback condition and the negative feedback condition. Two other designs were used to directly compare the effects of positive versus negative feedback, resulting in a 2 (Valence Feedback: positive/negative) X 2 (Task Domain: morality/competence) ANOVA and a 2 (Neutral Feedback: positive/negative condition) X 2 (Task Domain: morality/competence) ANOVA. These ANOVAs concerned a comparison between groups.

The analyses were carried out using the general linear model in SPM8. For each individual, contrast parameter images were computed and the resulting contrast images were submitted to second-level group analyses. Only effects of at least 10 continuous voxels that exceeded a False Discovery Rate (FDR) corrected threshold of $p < .05$ are reported.

Moreover, since we were interested in the –perhaps more subtle– difference between receiving feedback about morality or competence, we extracted parameter estimates from the regions of interest (ROI) that were identified in the whole brain analyses to explore the pattern of the activation across our conditions. We extracted the mean parameter estimate within each ROI for each condition, reducing the ROI to a single data point. This is a common approach in cognitive neuroscience which has two advantages: (1) it reduces the number of comparisons, and (2) collapsing across voxels within the region decreases noise (Poldrack, 2007). We focused specifically on the MPFC in the contrast positive versus negative feedback. However, activation in MPFC was part of a larger network (see Table 6.1). To isolate the activation cluster within the MPFC, we adjusted the threshold to $p < .01$ (FDR corrected, 10 continuous voxels, see Table 6.2). The ROI analysis

was used to gain functional specificity in the regions that were already a priori defined as regions of interest. This region was used to test the hypothesis that valenced feedback would be associated with differential activity in the morality versus competence condition. These regions were extracted using the Marsbar toolbox (Brett, Anton, Valabregue, & Poline, 2002) for SPM8.

Self-reported negative emotional response.

To examine participants' negative emotional response related to the moment they received their feedback, we used the same scales as described in Study 6.1: A scale measuring participants negative emotional response concerning their scores on morality (α = .90) and a scale measuring participants negative emotional response concerning their scores on competence (α = .89). These self-reports were administered after the scanning session.

Results

Behavioral data.

Since we asked participants to press a key whenever the fixation cross changed color (primarily to keep them attentive during the scanning session), we could test whether their response latencies differed between morality and competence trials. Indeed, a 2 (Feedback Type: positive/negative between-participants factor) x 2 (Task Domain: morality/competence within-participants factor) repeated measures ANOVA revealed a significant interaction effect; $F(1,37) = 9.54$, $p = .004$, $\eta^2_p = .21$. This indicated a significant reversal in the direction of the effects in the morality condition compared to the competence condition (see Figure 6.4). As a result, participants who received negative feedback responded significantly slower on morality ($M = 474.82$, $SD = 115.81$) than on competence trials ($M = 450.39$, $SD = 105.35$); $F(1,37) = 6.08$, $p = .02$, $\eta^2_p = .14$. In contrast, participants in the positive feedback condition responded somewhat more slowly on trials concerning competence ($M = 464.86$, $SD = 95.41$) than morality ($M = 444.24$, $SD = 66.77$); $F(1,37) = 3.71$, $p = .06$, $\eta^2_p = .09$.

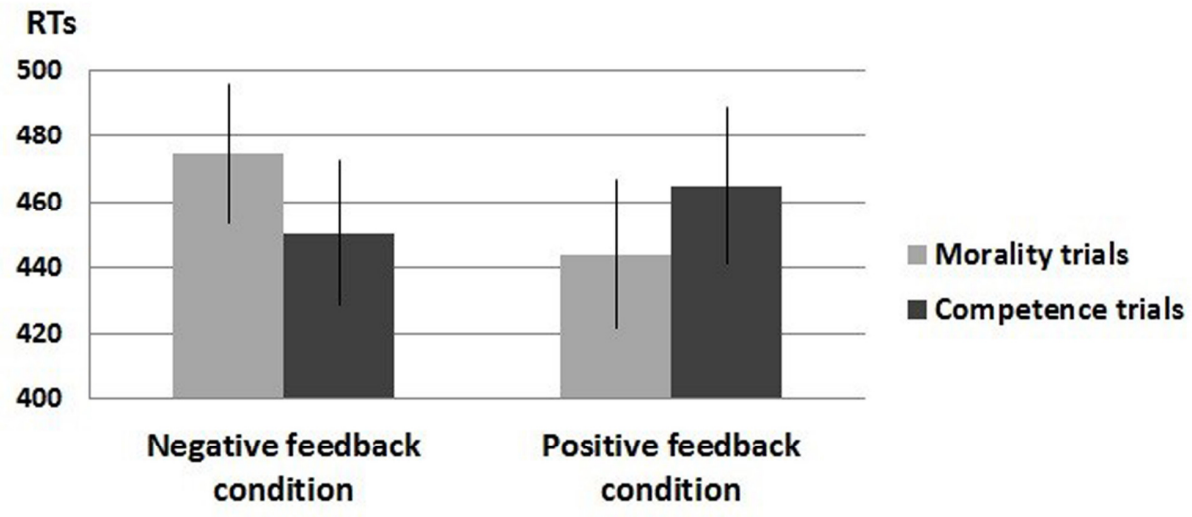

Figure 6.4. Interaction effect between reaction times (RTs, in milliseconds) on morality and competence trials: Whereas participants who received negative feedback responded more slowly on morality as compared to competence trials, participants who received positive feedback responded more slowly on competence as compared to morality trials.

fMRI data.

Whole brain level. To examine neural activation associated with receiving positive or negative and neutral feedback about one's morality and competence, we conducted four ANOVAs. First, we examined the effects of valenced and neutral feedback about morality and competence separately for the positive and negative feedback conditions. The results of these two 2 (Feedback: Valence/Neutral) X 2 (Task Domain: morality/competence) full factorial ANOVAs revealed no significant effects. Second, we examined neural differences between receiving positive versus negative feedback, by selecting only valenced trials. This 2 (Valenced Feedback: positive/negative) X 2 (Task Domain: morality/competence) ANOVA resulted in a main effect of valence (see Table 6.1): Activation in the amygdala, insula, bilateral inferior frontal gyrus, and ventral and dorsal MPFC was greater for participants who received positive feedback than for participants who received negative feedback. There was no main effect of Task Domain, nor an interaction effect. Third, we examined neural activation associated with receiving neutral feedback (i.e., only trials with neutral feedback were selected). This 2 (Neutral Feedback: positive/negative condition) X 2 (Task Domain:

morality/competence) full factorial ANOVA did not show any relevant significant activation (see Table 6.3).

Taken together, the contrast positive versus negative feedback resulted in activation in the expected brain network associated with processing self-relevant information. At the whole brain level the neural activation was not different for morality versus competence trials. In the next section, we describe the results from more fine grained ROI analyses, using the contrast positive > negative feedback as a functional localizer.

Regions of interest. To examine the difference between feedback related to morality and competence, we conducted ROIs analyses of the ventral MPFC (vMPFC), a target brain area showing increased activation for positive compared to negative feedback. Results revealed an interaction effect between feedback and task domain in the vMPFC (F[1,35] = 4.06, p = .05, η^2_p =.10). Consistent with our hypothesis that information concerning one's morality has a greater impact than information concerning one's competence, we found that the difference between positive and negative feedback was more pronounced for scores concerning morality; F(1,35) = 14.90, p < .001, η^2_p =.30, than for scores concerning competence; F(1,35) = 7.53, p = .01, η^2_p =.18 (see Figure 6.5). Moreover, within the positive feedback condition, activation in the vMPFC was greater when participants viewed their scores concerning morality as compared to competence; F(1,35) = 3.48, p = .07, η^2_p =.09. This difference was not significant in the negative feedback condition; F(1,35) < 1.

Self-reported negative emotional response. Results of a repeated measures ANOVA with task domain (morality/competence) as the repeated measure and valence (positive/negative feedback) as between-groups factor, supported our reasoning and were consistent with Study 6.1: We observed a significant interaction effect between task domain and valence; F(1,38) = 4.84, p = .03, η^2_p = .11. The relevant means and analysis of simple main effects confirmed that the difference between positive and negative feedback conditions in self-reported emotional response was more pronounced when participants received feedback regarding their morality; $M_{\text{difference}}$= 1.45, $S.E.$ = 0.22; F(1,38) = 44.24, p < .001, η^2_p = .54, rather than their competence; $M_{\text{difference}}$ = 0.82, $S.E.$ = 0.26; F(1,38) = 10.24, p =

.003, η^2_p = .21. Specifically, when participants had received negative feedback they indicated a more negative emotional response when the feedback was related to their morality (M = 4.14, $S.E.$ = 0.15) rather than their competence (M = 3.59, $S.E.$ = 0.18); $F(1,38)$ = 7.98, p = .01, η^2_p = .17. There was no difference between responses to positive feedback when comparing the morality with the competence domain (F < 1).

Table 6.1.

Brain regions revealed by the main effect of Valence in the 2 (Valenced feedback: positive/negative feedback) x 2 (Task Domain: morality/competence) ANOVA at whole brain level.

Anatomical Region	L/R	voxels	Z	MNI coordinates		
				x	y	z
Medial Orbital Prefrontal Cortex	R	51	3.43	30	47	-5
			3.27	36	29	-14
			3.12	27	53	1
Dorsal Medial Prefrontal Cortex	L	97	4.01	-12	26	34
			3.26	-5	38	31
Dorsal Lateral Prefrontal Cortex	L	164	4.53	-36	35	13
			3.09	-45	14	19
			2.72	-57	11	22
Superior Frontal Gyrus	L	27	3.34	-18	26	55
Supplementary Motor Area	R	16	2.85	3	8	61
Middle Temporal Gyrus	R	11	2.74	48	-58	19
ParaHippocampal Gyrus	L	18	3.11	-24	-37	-8
Calcarine/Linual Gyrus	R	5575	5.66	15	-88	10
			5.26	18	-55	-2
			5.07	15	-64	16
Middle Occipital Gyrus	L	44	3.73	-30	-88	19
			3.47	-27	-91	10
			2.74	-33	-73	28

MNI coordinates for main effects, peak voxels reported at p < .05, FDR corrected, at least 10 contiguous voxels (voxels size was 3.0 x 3.0 x 3.0 mm).

Table 6.2.

Brain regions revealed by the main effect of Valence in the 2 (Valenced feedback: positive/negative feedback) x 2 (Task Domain: morality/competence) ANOVA at whole brain level.

Anatomical Region	L/R	voxels	Z	MNI coordinates		
				x	y	z
Ventral Medial Prefrontal Cortex	R	88	4.34	0	59	-2
			4.29	-12	59	10
			4.09	9	59	-2
Dorsal Lateral Prefrontal Cortex	L	27	4.53	-36	35	13
	R	12	3.97	48	20	28
Rolandic Operculum/Precentral Gyrus	R	108	4.74	51	-13	19
			4.37	48	5	37
Pre-/Postcentral Gyrus	L	19	4.40	-54	2	40
			3.81	-51	-10	37
Superior Temporal Gyrus	R	23	4.17	57	-4	4
Middle Temporal Gyrus	L	19	3.91	-45	-67	19
Superior Parietal Lobule	L	72	4.72	-24	-58	55
Superior Parietal Lobule / Cuneus	L	72	4.72	-24	-58	55
	R	130	4.77	12	-85	31
			4.51	15	-64	55
			4.29	18	-58	46
Precuneus	L	16	3.78	-6	-58	37
			3.49	-12	-58	31
Calcarine/Linual Gyrus	R	379	5.66	15	-88	10
			5.26	18	-55	-2
			5.07	15	-64	16
Middle Occipital Gyrus	R	16	4.46	30	-79	31
Insula	L	52	4.12	-33	-16	16
			3.82	-24	-19	19
	R	13	3.93	39	-28	22
Amygdala	R	10	4.13	30	2	-14
Hippocampus	R	21	3.96	24	-34	-5

MNI coordinates for main effects, peak voxels reported at $p < .01$, FDR corrected, at least 10 contiguous voxels (voxels size was 3.0 x 3.0 x 3.0 mm).

Table 6.3.

Brain regions revealed by the main effect of Valence in the 2 (Neutral feedback: positive/negative condition) x 2 (Task Domain: morality/competence) ANOVA at whole brain level.

Anatomical Region	L/R	voxels	Z	MNI coordinates		
				x	y	z
Superior Parietal Lobule	R	21	4.52	18	-61	55
			3.95	18	-58	46
			3.68	21	-55	43
Calcarine Gyrus (Occipital Lobe)	R	47	4.30	24	-61	19
			4.22	15	-64	16

MNI coordinates for main effects, peak voxels reported at $p < .05$, FDR corrected, at least 10 contiguous voxels (voxels size was 3.0 x 3.0 x 3.0 mm).

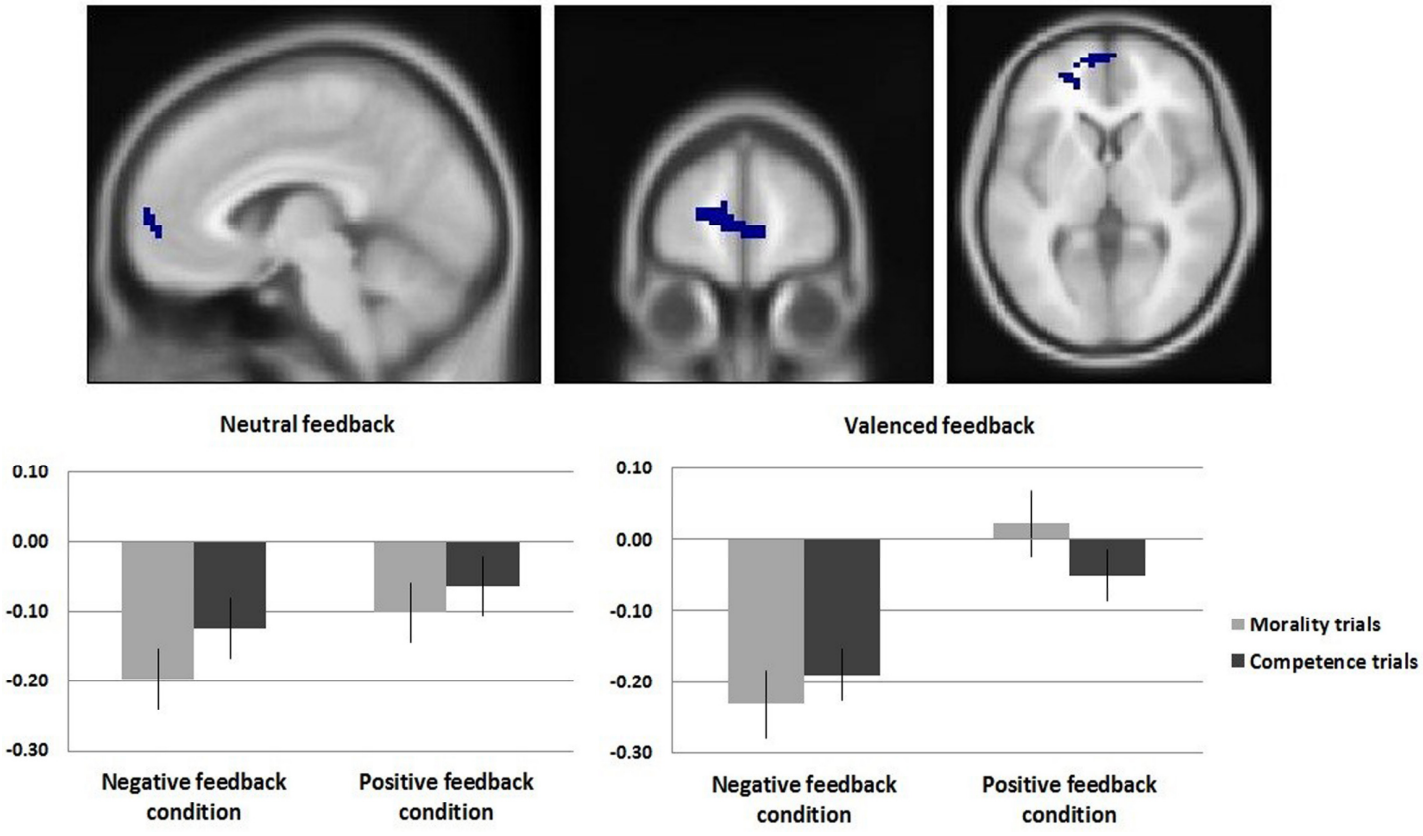

Figure 6.5. Activation in the ventromedial prefrontal cortex (ROI cluster based on a peak voxel, MNI coordinates: x = 0, y = 59, z = -2; $p < .01$, FDR corrected, $p < .01$, at least 10 continuous voxels) revealing the significant interaction between feedback and task domain on valenced feedback trials. There were no effects on neutral feedback trials.

General Discussion

The aim of the present research was to compare the impact of receiving different types of self-relevant information. Specifically, we compared behavioral, self-reported, skin conductance and fMRI responses to information regarding an individuals' own morality and competence. Previous research revealed that when receiving information about another person's morality, negative behaviors are perceived as more informative than positive behaviors. Conversely, in the competence domain, positive information is perceived as more informative than negative information (Skowronski & Carlston, 1987). Importantly, however, this differential diagnosticity has been demonstrated when people form an impression of others. Thus, it is as yet unclear whether a similar asymmetry in the perceived importance of positive and negative information regarding competence and morality is also evident when people process information related to the self.

We examined this in the present research, by confronting participants with information either attesting to or undermining their moral and competent self by giving them positive or negative feedback about their performance on a task that was supposedly indicative of both domains. After having received the feedback, we asked participants to recall their affective responses (i.e., positive and negative emotions) related to the moment of feedback. Additionally, we assessed participants' physiological arousal by assessing their skin conductance levels while they received their feedback (in Study 6.1) and (in Study 6.2) we used fMRI to examine how activation in the neural network involved in processing self-relevant information, was associated with receiving the feedback.

Participants self-reported emotions gave insight into how people reflect upon the information they received about their moral and competent self and thus whether this self-reflection mirrors the asymmetry that has been observed in impression formation of others. The evidence obtained provided partial support for our reasoning regarding the differential diagnosticity of (im)moral and (in)competent information about the self. That is, compared to information concerning competence, information concerning morality had a greater impact upon participants' self-reported emotional response. Especially participants who had received negative feedback reported increased negative affect when the

feedback concerned their morality rather than their competence. These findings extend research about the importance of morality over competence for people's personal and social identity (e.g., Leach et al., 2007; Ellemers et al., 2008).

Interestingly, the results of our (neuro)physiological measures offered additional support for the pattern of differential diagnosticity of (im)moral and (in)competent behaviors found in impression formation research. That is, results of analyses of skin conductance responses revealed that physiological arousal was increased when participants received feedback about their morality as compared to their competence, and this was the case in particular when this feedback had a negative content. (Negative) information about one's own morality thus seemed to be more impactful than information concerning one's competence. These findings thus extend prior research which established the explicit motivation to be (perceived as) moral (e.g., Leach et al., 2007; Ellemers et al., 2008) as they reveal that automatic affective responses are increased when people are confronted with information that calls their morality into question.

In addition, results of the fMRI experiment showed that positive (rather than negative) feedback was associated with greater activation in the amygdala, insula and MPFC. The MPFC has previously been associated with the processing of self-relevant information (e.g., see Abraham, 2013; Moran et al., 2006; Northoff & Bermpohl, 2004; Schmitz & Johnson, 2007). The relative increase in activation in this region for participants who received positive feedback (as compared to participants who received negative feedback) is in line with research showing that people are positively biased when they receive self-relevant information. Specifically, people tend to think they are better than average (especially when the other is a non-specified average student, like in our study; Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995), and expect to receive positive rather than negative feedback in social interactions (Hepper, Hart, Gregg, & Sedikides, 2011). Moreover, prior research has established that positively biased feedback processing is associated with activation in the MPFC (Korn, Prehn, Park, Walter, & Heekeren, 2012). Extending this prior research, our findings thus reveal that positive information concerning one's own behavior is processed as more self-relevant than negative information concerning one's behavior. Moreover, our

results revealed that participants showed more activation in the ventral MPFC (vMPFC) when they received positive feedback concerning their morality as compared to their competence. In line with research suggesting that the vMPFC is associated with the detection and labelling of information relevant to the self (Van der Meer et al., 2010), these findings thus suggest that people detect of the confirmation of one's morality as more self-relevant than confirmation of one's competence.

The findings concerning the impact of negative feedback on affective responses and arousal complement the observed effects of positive feedback in the fMRI results. That is, the skin conductance data in combination with the self-reports suggest that people are emotionally moved by negative feedback concerning their own moral behavior. Additionally, they process positive feedback concerning their own moral behavior as more self-relevant. Across the board, people thus seem more likely to attend and respond to information regarding their morality rather than their competence, which suggests that this process is more complex than the process of impression formation of others: Consistent with impression formation, negative information about the self also has a greater impact when it concerns one's own morality as compared to competence. However, people also seem to be more attentive to positive information concerning their own morality and what this means for their self-view, than that they are focused on possible implications of negative information concerning their own morality. In other words, people are thus particularly attentive to moral information that may help establish a positive self-view. Again, such positive information is most relevant when it concerns one's morality rather than one's competence. At the same time, although people seem to attend less to negative information concerning the self, being confronted with such negative information induces increased arousal and negative emotion. Unfortunately, we cannot directly relate the findings concerning the skin conductance to the fMRI data since we assessed these measures in two separate studies. In order to examine this relation more directly, a measure of skin conductance should be taken *while* participants are being scanned. Nevertheless, different from how we respond to information about others – when negative information is seen as more indicative of another person's morality, our present

observations suggest that we seem to perceive positive information as most relevant to ourselves, especially when this indicates and confirms our moral identity.

Acknowledgements

We thank Thijs Schrama and Maureen Meekel for their assistance with processing the skin conductance data; Maryke Hofman and Lotte van Dillen for their help with the fMRI data collection; and Ilya Veer and Mischa de Rover for their advice concerning the paradigm for the fMRI study.

Chapter 7

General Discussion

In this dissertation, I examined three different research questions. In Part I, I tested whether people tend to act in ways that are considered moral. In Part II, I addressed the question how important is it for people to be perceived as moral by others. In Part III, I examined how much people care about whether or not they succeed in behaving according to their moral values. Additionally, I aimed to unravel the cognitive processes associated with these motivations. In this final chapter, I will discuss the conclusions that can be drawn from the research reported in the previous five chapters. First, I will review the behavioral findings observed in the three different parts of the dissertation. Then I will elaborate on evidence revealing the underlying processes associated with these behavioral results.

Behavioral Findings

Part I: Moral Concerns Cause Inhibition of Intergroup Bias

Previous research has revealed that people explicitly report that they think it is more important to be perceived as moral than as competent (Ellemers, Pagliaro, Barreto, & Leach, 2008). One aim of the current dissertation was to examine whether people not only explicitly report this motivation, but actually tend to behave more according to their moral than competence values. To assess this, I presented native Dutch, non-Muslim, research participants with an implicit association test (IAT). This test is a measure of one's (implicit) prejudice towards a particular outgroup –in my research these were Muslim women. I framed this test as being able to show how moral or how competent people are. Specifically, participants were either informed that "this task can give an indication about your moral values concerning egalitarianism and discrimination", or that "this task can give an indication about your ability to learn new tasks and to quickly process new information". I thus examined whether the implicit bias people showed in their task behavior would be reduced to a greater extent when they were motivated to be moral than when they were motivated to show their competence. Results in Chapters 2, 4 and 5 (Studies, 2.1, 2.2, 4.1, and 5.1) revealed that people indeed are more motivated by their moral than their competence values. In case of an emphasis on the moral (as compared to the competence) test implications, participants were more likely to control their negative bias against Muslim women.

In this dissertation research, no specific norms were made salient when the moral implications of the task were emphasized, nor were participants explicitly informed on how they might avoid displaying bias while working on the IAT. In the moral motivation condition, participants only read that the test could give an indication of the value they attached to egalitarianism vs. discrimination. Furthermore, participants' performance was assessed when they performed the test in private and anonymously. The findings obtained with this procedure extend previous research as they make it possible to exclude a number of alternative explanations relating to self-presentation and displays of socially desirable response patterns. Thus, the data reported here reveal that people act upon their own moral values, presumably because this is important for how they see themselves.

Part II: Moral Motivation is Affected by the Social Context

In the second part of this dissertation, I examined whether it is also important for people to be perceived as moral by others. To examine this I introduced a procedure that led participants to believe that their performance on the IAT was monitored by another individual present in the lab. My findings reveal that people are particularly motivated to act according to their moral values in the presence of people who belong to the same group as they do (i.e., ingroup members). In my research, these groups were created according to very minimal criteria (i.e., a minimal group paradigm; Tajfel, 1970). Before participants started with the IAT, they completed a questionnaire that ostensibly assessed their personality type. After that, they were told that they were coupled with their evaluator based on both their questionnaire scores. It was explained that when the evaluator was assessed to have the same personality type as the participant, they shared this particular group membership. In other words, this made them ingroup members. When the evaluator was assessed to have another personality type, s/he differed from the participant. This distinction in their personality types thus made the evaluator an outgroup member. This type of paradigm allowed me to exclude the possibility that alternative concerns (such as prior liking, familiarity or value similarity) might induce participants' responses to different evaluators. Thus, I was able to establish that people are more motivated to act morally in front of others who are relevant to the self. In real life, these self-relevant others might include people with the same

nationality, gender, religion, or occupation. The current findings thus extend previous research that demonstrated that people explicitly report the importance of being seen as a moral ingroup member (Leach, Ellemers, & Barreto, 2007) by showing that they actually are more likely to act in accordance with their moral values when their behavior is monitored by an ingroup member.

One could wonder how meaningful the situation created in this experiment is, as a minimal group paradigm is unrelated to the intergroup associations examined with the IAT. That is, in the IAT participants are asked to make associations between non-Muslims and Muslims and pictures of positive and negative scenes. Group memberships based on personality type may thus be seen as irrelevant to the task. However, exactly because of the use of a minimal group paradigm, I was able to reveal the importance of being perceived as moral for people's social identity. Introducing two experimentally created groups, which had no meaning or known moral values outside of the laboratory, was sufficient to increase people's motivation to appear moral towards an ingroup rather than an outgroup member. This finding can thus not be attributed to factors other than the categorization allegedly based on personality types introduced in the experiment.

Nevertheless, the importance of being perceived as moral by self-relevant others may go beyond a shared minimal group membership. In fact, introducing a group membership that does interfere with the intergroup associations made in the IAT may reveal additional motivations to adhere to moral norms. In Chapter 5 I accordingly established that being perceived as unprejudiced is even more important when a representative of the social target group is present. That is, when participants thought that their performance was monitored by a Muslim woman, they inhibited their bias against Muslims to an even greater extent than in the presence of a minimal ingroup member. Although this finding is consistent with previous research on intergroup bias (e.g., Lowery, Hardin, & Sinclair, 2001; Richeson & Ambady, 2003), the current results extend this research in an important way. The way my study was set up allowed me to show that the moral implications of one's behavior can be emphasized in many different ways – that all can be effective. The results of Study 5.1 show that simply mentioning the moral implications of the task affected people's implicit prejudice in a similar vein as did

the actual presence of a Muslim evaluator. That is, my research demonstrated that participants inhibited their bias to a similar extent when they were being monitored by a non-Muslim woman during a task of which the moral implications were emphasized as when they were being monitored by a Muslim woman. Taken together, these findings thus reveal that signaling the moral implications of one's performance can instigate moral behavior to a similar extent as explicitly confronting people with others who depend on them for moral treatment (i.e., Muslims).

The Underlying Processes

Parts I and II

Apart from showing behavioral effects of emphasizing the moral implications of one's behavior, I also examined the cognitive processes underlying people's motivation to be and to appear moral. More specifically, by applying measures borrowed from the field of neuroscience, I was able to show how focusing on people's morality changes their attention to ingroup and outgroup members, as well as the degree to which they monitor their own moral behavior.

Moral motivation changes people's focus of attention.

In Studies 2.2, 3, and 4.2, I examined brain activation associated with the motivation to be moral using event-related brain potentials derived from EEG (i.e., ERPs, derived from activation recorded at the scalp) and functional MRI (i.e., to localize activation in the brain) while they were performing moral behavior. Results showed that emphasizing the moral implications of people's behavior causes them to increase their attention towards the faces of the different group members presented in the IAT. People thus attended more to the difference between ingroup and outgroup members when they were motivated to approach this task in a moral way compared to when they were concerned with being competent at the task. At first sight this increased attention to group membership may seem to contradict moral intentions. That is, performing in line with moral values –not revealing intergroup bias– might also be expected to result in less differentiation between groups evident in increased similarity of cognitive responses when looking at members of ingroups and outgroups. Nevertheless, while participants were more inclined to attend to the group membership of the target stimuli under moral task

instructions, we found in Study 2.2 that people were more able to respond in an unbiased way. This combination of effects seems to suggest that the increased social categorization of ingroup as distinct from outgroup members was needed in order for participants to inhibit their bias against the outgroup and thus to adhere to their moral values. This explanation is in line with the notion that in order to deal with one's upcoming thoughts, these must first be recognized and accepted (Wegner, 2011). Likewise, in order to suppress the tendency to reveal bias, one must first acknowledge the difference between group members.

The investigation of the cognitive processes underlying moral motivation also extend our findings on the behavioral measures revealing the importance people attach to being perceived as moral by their ingroup members in particular. That is, complementing the behavioral effects observed in Study 4.1, the results of Study 4.2 revealed that participants' increased cognitive attention to the ingroup and outgroup faces when the implications of the test were formulated in terms of their moral values, only emerged when they were evaluated by someone of their own (minimal) ingroup, and not when they were being monitored by a member of another (out)group. In other words, the adjusted cognitive approach towards the task –arguably to make it possible to adhere to moral group norms– was especially apparent in an intragroup context.

Moral motivation enhances response-monitoring.

Besides the increased perceptual attention to the difference between faces of ingroup and outgroup members, participants to whom the moral (rather than the competence) implications of the task were emphasized also showed enhanced error-monitoring. That is, when participants were motivated to show their morality, they paid more (automatic) attention to their *responses* than when they wanted to show their competence. Consistent with previous ERP findings (e.g., Gehring, Goss, Coles, Meyer, & Donchin, 1993), the error-related negativity (ERN) modulation was evident when participants made incorrect (as compared to correct) responses. Importantly, this enhanced response to errors was greater when the moral rather than competence implications of the task were emphasized. Previous research has revealed that increases in the ERN are associated with how important a good task performance is to people, which is indicated by the extent to which

they care about making errors on the task (Hajcak, Moser, Yeung, & Simons, 2005). The enhanced ERN modulation in case of an emphasis on one's moral values thus implies that people are more concerned about making mistakes when the task supposedly indicates their morality than when it 'merely' indicates their competence. Additionally, these findings suggest that the motivation to be moral can in part be explained by an increased concern about not appearing moral. In comparison, the prospect of appearing incompetent seems to be less distressing.

Importantly, the effect concerning participants' error-monitoring also depended on the social context. Results of Study 4.2 showed that the emphasis on moral implications in combination with being monitored by an ingroup member increased response-monitoring on both incorrect *as well as* correct responses. Thus, when people are evaluated by another ingroup member, they show a general increase of attention towards their own moral behavior. The ERN findings in Part I thus reveal that people are primarily concerned with making mistakes that can be perceived as an indication of immoral behavior. In addition, the results in Part II show that when people show their moral behavior to their fellow group members, it seems equally important to detect any mistakes as it is to monitor their correct responses.

Moral motivation increases detection of task-relevant characteristics.

The emphasis on the moral implications of the task and being monitored by an ingroup member also affected participants' detection of the different types of trials in the IAT. The IAT consists of incongruent and congruent trials. As participants who took part in the research described in this dissertation were non-Muslim, the congruent trials consisted of associating faces of non-Muslim women (i.e., ingroup members) with pictures of positive scenes, and faces of Muslim women (i.e., outgroup members) with pictures of negative scenes. By contrast, the incongruent trials consisted of associating outgroup members with positive pictures and ingroup members with negative pictures. Previous ERP research has shown that the detection of the difference between such congruent versus incongruent trials (i.e., 'conflict-monitoring') is visible in the N450 modulation, which is typically larger for incongruent than congruent IAT-trials (e.g., Williams & Themanson, 2011). Results of Studies 2.2 and 4.2 showed that the N450

modulation was increased in case of an emphasis on morality and when an ingroup member was evaluating participants' performance. The detection of the different types of IAT trials was thus enhanced under these circumstances. A possible explanation for this finding may be that the moral implications of the test, and the presence of an ingroup member, made the meaning of the difference between congruent versus incongruent trials more evident. It suggests that participants may have understood that the ease with which they responded on congruent as compared to incongruent trials was related to possible signs of prejudice. These participants may have realized that the relatively easy part of the task consisted of associating the outgroup with negativity and the ingroup with positivity. And that the relatively difficult part meant associating the outgroup with positivity and the ingroup with negativity. In contrast, participants who thought the task was indicative of their competence may only have noticed the difference in the level of difficulty between the two types of trials, without taking a notion of the social meaning of the associations they were asked to make.

Overall, the findings of Parts I and II are important as they extend prior research that used self-reports (e.g., Leach et al., 2007; Ellemers et al., 2008) as well as our own research showing actual moral behavior on an IAT to examine the importance of being moral. By incorporating the examination of unconscious cognitive processes with ERP measures, the current findings reveal *how* people's motivation to be (perceived as) moral leads to more moral behavior. Results concerning the underlying cognitive processes reveal that moral concerns affect how people perform the task and to what kind of aspects they pay attention during the task (i.e. "Is this person a Muslim or non-Muslim?"; "Is this particular trial more or less difficult?" and "Am I succeeding in being unbiased?"), affecting their actual moral behavior (in this case implicit bias against Muslims).

Part III: People Show a Positivity Bias Concerning Their Own Morality

Overall, the behavioral, ERP and fMRI results of the first two parts of this dissertation indicated that emphasizing the moral implications of one's behavior (either while being evaluated by an ingroup member or in private) causes people to become more vigilant during their performance on a test of implicit prejudice. The findings also seem to suggest that adherence to moral norms is equally important as

it is to avoid committing moral transgressions. A possible explanation could be that the motivation to be moral is accompanied by a fear to appear immoral, whereas the possibility of appearing incompetent may be less distressing. That is, in our mind even competent people may sometimes do incompetent things, but people who do something immoral once are unlikely to be seen as moral persons. In Part III of this dissertation, I therefore examined how much people care about whether they succeed or fail in behaving according to their moral values – compared to how much they care about their success or failure in the competence domain.

In Chapter 6, I assessed people's affective and cognitive responses after and while they received information about their own moral and competent behavior. Participants first performed a task (the IAT), but in contrast to the procedure in previous chapters, this task was said to be indicative of their moral values *as well as* their competence. Thereafter, they were either informed that they had performed above (positive feedback) or below (negative feedback) average on the moral and competence dimensions of the task. This allowed me to directly compare how positive versus negative feedback concerning one's own moral and competent behavior impacted upon people's state of mind and emotional well-being.

Results of Study 6.1 revealed that people feel bad when they are confronted with information indicating that they are not that moral as compared to others. Such information causes increased levels of physical arousal (measured using skin conductance responses) and people also report to experience more intense negative emotions. Crucial to my predictions, receiving information that one is less moral than others makes people feel worse than receiving information indicating that they are less competent than others. These findings thus confirm the notion that people care more about whether they succeed in behaving according to their moral values rather than behaving competently.

Additionally, results of the fMRI study in Chapter 6 seem to suggest that when people receive positive information indicating that they are more moral compared to others, they perceive this information as highly relevant to their self-concept. Previous neuroimaging studies showed that activation in the (ventral) medial part of the prefrontal cortex (vMPFC) is associated with ascribing personal characteristics or behaviors to the self (e.g., Van der Meer, Costafreda, Aleman, &

David, 2010). In line with the notion that people want to be moral, I thus examined whether viewing information indicative of one's own moral behavior is associated with activation in the vMPFC. Indeed, results of Study 6.2 showed that activation in the vMPFC was greater when participants viewed feedback about their moral behavior as compared to their competent behavior. Interestingly, this was only the case when this feedback consisted of positive information. The results of Chapter 6 thus reveal that people seem to perceive positive indicators of their moral behavior as particularly self-relevant. A tentative explanation for this finding could be that participants in this study protected themselves from negative feedback by processing it as relatively less self-relevant. This is in line with my observation that the confrontation with negative indicators of one's own morality has a highly negative impact upon people's emotional well-being. Hence, discarding such information as less self-relevant might be part of a self-protective strategy to cope with such threatening information. Taken together, the findings thus confirm how much people care about succeeding in behaving in line with their moral values, and how they respond to information that may indicate this.

The Added Value of Different Research Methods

In this dissertation, I addressed three research questions related to people's motivation to be (perceived as) a moral individual and group member. A primary aim of the dissertation was to examine the underlying processes associated with this motivation. I thus combined behavioral observations with psychophysiological and neuroscientific research tools throughout the empirical chapters to go beyond observing *what* people do, and examined *how* and *why* they do this in terms of specific underlying processes.

The behavioral task used in the empirical chapters provided reaction times and error rates. It showed us that people inhibit their bias against Muslim women by slowing down their responses on prejudice-congruent trials. This measure thus revealed *what* people do, but it remains unclear *how* they are able to do this. Likewise, self-report measures are often administered after a particular behavior is displayed. Such measures rely upon the ability and willingness of research participants to report on their psychological state while performing the task, and are sensitive to social desirability – which obviously is a significant factor in

research concerning the motivation to appear moral. This is why it was important for me to assess the psychophysiological and neuroscientific measures *online*, that is, while participants were actually performing the task. The neural and physiological reactions I assessed occur unconsciously and are less sensitive to the intention to respond in socially desirable ways.

Using ERPs, I was able to disentangle different cognitive processes associated with the control of prejudice. In this way, I revealed three different mechanisms that help participants inhibit their (behavioral) bias against Muslims. They did this: (a) By (unconsciously) increasing their perceptual attention to categorize target faces as Muslim versus non-Muslim women; (b) by distinguishing between prejudice-congruent and –incongruent trials; and (c) by monitoring their responses during the task. Recording skin conductance responses allowed me to reveal that receiving information about people's own moral behavior causes instant automatic arousal that is different from how they respond to information concerning their competence. These findings thus underscored participants' explicit reports of their negative affective states. Furthermore, based on fMRI-results –and particularly the knowledge of the functional properties of activation in the ventral medial part of the prefrontal cortex– I have suggested that people perceive positive information indicating their morality as particularly relevant to their self-concept.

To give a concrete example of the added value of the different research methods combined in the current dissertation, let's consider the findings in Chapters 2 and 4. Here, I discovered that an emphasis on the moral implications of one's behavior affects people's approach towards a task. Using a behavioral measure of implicit prejudice, I showed that non-Muslim participants inhibited their negative bias against Muslims when they were told that the test could assess their moral values concerning egalitarianism (as compared to how competent they are; Studies 2.1 and 4.1). The weaker negative bias when the moral test implications were stressed was caused by a smaller difference in response times between incongruent and congruent trials. This means that when morality was emphasized, non-Muslim participants responded equally slowly to congruent trials (associating non-Muslims with positivity and Muslims with negativity), as to incongruent trials (associating non-Muslims with negativity and Muslims with positivity). In this

sense, people thus made less of a distinction between their associations with ingroup and outgroup targets, and this is what resulted in the reduction of bias. However, interestingly, examination of brain activation during task performance revealed a significant difference between viewing pictures of ingroup and outgroup targets. That is, ERP modulations associated with differentiating between viewing in- and outgroup targets were increased rather than decreased (Studies 2.2 and 4.2). Additionally, activation in the occipital face area was greater for viewing faces of ingroup compared to outgroup targets when morality was emphasized (Study 3).

At first glance, the behavioral and neuroscientific research findings thus seem to be contradictory. On a behavioral level, emphasizing morality resulted in more *equal* responses to members of different groups, whereas emphasizing morality actually increased *differentiation* between groups at the neural level. However, it is important to understand that both measures assessed different cognitive processes which occur at different stages in the process. That is, behavioral bias was estimated based on reaction times and the accuracy of responses on all trials within the task. This includes trials with pictures of faces of in- and outgroup targets, and on trials with pictures of positive and negative scenes. By contrast, perceptual attention was assessed from early visual processing of faces alone, irrespective of the response given on these types of trials. This may imply that the salience of morality *increased* people's perceptual *attention* towards the faces of in- and outgroup members, and this is what enabled participants to *behaviorally* respond with *decreased* bias, in line with their moral values. It also suggests that participants actually attended differently to specific task stimuli when its moral implications were emphasized, rather than merely correcting their behavioral responses to these stimuli. This combination of observations thus suggests that the adjustment in participants' behavior may at least in part depend on early cognitive processes that are crucial for preparing these responses.

Another example of the added value of the new approach I followed in the current dissertation concerns the examination of the behavioral IAT effect. Across the different studies reported in Parts I and II, I found the same effect of emphasizing the moral implications of participants' IAT performance. Participants to whom the moral implications of the task were emphasized (and whose

performance was monitored by an ingroup member) showed a relatively weak negative bias against Muslims. Interpreting this finding based only on the strength of the bias does however not inform us about how this reduction in bias is achieved. For example, the emphasis on morality may have caused participants to develop stronger positive associations with Muslims. On the other hand, it could also have helped them to control their negative associations with the Muslim targets. Either way this might have increased perceived equality between the two target groups, resulting in the smaller negative bias against Muslims that was found. Nevertheless, I set out to examine which process actually resulted in these behavioral findings.

Previous research concerning the malleability of implicit bias can be seen to represent two distinct approaches. There are studies (such as the ones described in this dissertation) that examine what kind of motives or contextual factors affect displays of prejudice (i.e., the dependent measure of social bias such as the IAT effect). There are also studies that focus on how people's performance on such measures of prejudice can be influenced (e.g., Fazio & Olson, 2003; Olson & Fazio, 2003). In this second type of research different models have been introduced to examine responses on reaction times and error rates to disentangle the processes underlying automatic evaluations and control. Examples are the process-dissociation model (e.g., Jacoby, 2001; Payne, 2001); the Quad-model (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005); the diffusion-model analysis (Klauer, Voss, Schmitz, & Teige-Mocigemba, 2007); and the ReAL model (Meissner & Rothermund, 2013). In my research, I did not follow any of these particular models, but adopted a more general strategy to examine the underlying processes associated with the reduced bias in case of an emphasis on moral values. I more closely examined participants' response times on *correct* congruent and incongruent IAT trials, to be able to distinguish between two different routes towards bias reduction.

In theory, bias measured using an IAT can be diminished in two ways. Either by reducing response times on incongruent trials, or by increasing response times on congruent trials. The first strategy implies that participants try to respond faster when they are asked to associate outgroup members with positive attributes and

ingroup members with negative attributes. The second strategy implies that participants slow down their responses when they are asked to associate outgroup members with negative attributes and ingroup members with positive attributes. As described in each of the relevant chapters, the current results showed that the emphasis on morality (and being evaluated by an ingroup member) caused participants primarily to *slow down* their responses on prejudice-*congruent* trials. This suggests that stressing the moral implications of their performance made the meaning of congruent trials –as potentially revealing biased associations– more salient. That is, under moral task instructions participants were more likely to realize that congruent trials would reflect the ease with which they could associate Muslims with negative attributes and non-Muslims with positive attributes. My approach to examine how the emphasis on moral values affects people's bias towards Muslims thus revealed that this may have led them to slow down and overthink these prepotent responses, to be able to act in line with (self-relevant) moral values.

The Challenges of Different Research Methods

In my research, I set out to combine procedures and measures that had been developed in different research traditions, to examine distinct research questions. Combining different approaches in this way certainly had an added value for my research and the conclusions I was able to draw. Nevertheless, I also had to face several complications relating to adjustments I had to make to experimental designs and standard procedures, to adapt the IAT for use of different neuroscientific research methods.

In Study 2.2, I had a clear hypothesis about how the ERN modulation would be affected by the emphasis on the moral implications of the task. But in order to reliably estimate the ERN, a sufficient number of errors is needed. I thus doubled the amount of trials in the IAT in order to allow participants to reveal more mistakes during their task performance. Although prolonging the IAT did result in the intended increase in errors, it also caused a learning effect: After so many trials, regardless of condition, all participants responded in the same way to all types of IAT trials. This adaptation of the task to enable examination of the ERN modulation thus reduced the difference in performance depending on whether

moral or competence task implications had been emphasized. As a result, the behavioral effect I had found in Study 2.1 was less clearly visible in Study 2.2.

In Chapter 3, I faced a similar problem, when I did not find an effect of emphasizing moral test implications on the behavioral data at all. In this fMRI study, I used an event-related design to be able to detect brain activation associated with the presentation of the different types of stimuli. To be able to separate responses to different trials, this design requires that there is a certain waiting period in between each of the trials. The time delay between trials, required to reliably assess fMRI responses, slowed down the overall pace of the IAT, and may have helped participants to prepare for and focus their attention for each upcoming trial. As a consequence, when using this procedure, the response times of participants who read the moral implications of the test were similar to the response times of participants who read the implications concerning their competence. This aspect of the task procedure may explain why I was unable to demonstrate the previously presented behavioral effects of the emphasis on one's moral values, in this particular study.

Finally, the behavioral effect of reduced bias in case of the morality frame in combination with evaluation by an ingroup member (observed in Study 4.1) did not emerge in (ERP) Study 4.2 in which I examined the effects of morality framing and presence of ingroup versus outgroup members. In retrospect, this may be attributed to the limited response window we offered to participants. As was the case in Study 2.2, I adapted the IAT procedure in Study 4.2, because I needed enough erroneous responses to reliably estimate the ERN. A pilot study had however uncovered that participants responded more accurately as well as faster when their performance was being monitored. Thus, in addition to doubling the number of trials like I did in Study 2.2, in Study 4.2 I also tried to induce participants to make a sufficient number of errors by reducing the time available to respond on each trial. In Studies 2.1, 2.2, and 4.1, the decrease in behavioral bias against Muslims was associated with participants' slowed down responses on congruent trials. Slowing down was however no longer possible in Study 4.2 because of the limited response window. This might explain why no evidence of reduced behavioral bias was found here. Nevertheless, the ERP measures

confirmed that the underlying cognitive processes were affected when the moral implications of the task were emphasized and when participants were monitored by an ingroup member.

Unfortunately, such difficulties are inherent to the choice of combining different research methods –while using the same behavioral paradigm–, to obtain triangular evidence as a way to examine complex psychological questions (see also Scheepers, Ellemers, & Derks, 2013). However, importantly, the fact that these adaptations had to be made and affected the results also extended current insights in the processes underlying the influence of (moral) motivation on IAT performance. For example, in Study 5.2, I also extended the number of trials included in the IAT. This time, I examined whether increased exposure to an apparent outgroup member (i.e., a woman with a headscarf) who was presented as a partial ingroup member on another dimension (i.e., in terms of her personality type) might increase positive associations with Muslim women. As in Study 2.2 – where the IAT effect was extinguished over time– the prolonged IAT caused a learning effect once again. But this time extending the number of IAT trials enabled participants to *develop new associations*, by learning to combine positive stimuli with the outgroup target. The evidence that it is possible for participants to do this is important beyond its methodological implications, as it offers scope for developing very practical and concrete strategies that may help reduce the emergence of implicit negative biases by learning to make new associations (see also Kawakami, Dovidio, Moll, Hermsen, & Russin, 2000).

Extending Previous Literature

The research presented in this dissertation extends existing insights in many ways. Importantly, this research is the first to show that morality not only induces people to *say* that they want to behave in a certain way, but that it actually motivates them to change their *behavior*. Extending previous research that focused on people's explicitly reported moral motivation and stated preferences in hypothetical moral dilemmas, I was able to reveal that people adjust their implicit behavior (i.e., their IAT performance) in line with their moral values when the moral implications of that task are made salient.

The findings reported in the current dissertation also have important implications for neuroscientific research. Previous research has examined the brain regions involved in the ability to behave in line with moral standards by studying patients with brain lesions who –as a result– exhibit immoral behavior or psychopathological characteristics (e.g., Anderson, Bechara, Damasio, Tranel, & Damasio, 1999; for a review see also Moll, Zahn, De Oliveira-Souza, Krueger, & Grafman, 2005). In other cognitive research there has been a focus on moral decision making. There, it is examined which parts of the brain people need (i.e., need to be activated) to *consider* what one would do in a hypothetical situation. The findings in this dissertation thus extend those insights by revealing the cognitive processes involved when healthy participants *behave* (i.e., perform) in line with their own moral values. Moreover, by studying the brain regions involved in the motivation to appear unprejudiced in a situation that resembles common interactions (when first viewing faces of people representing different social groups), I was able to examined a kind of moral motivation that is part of social interactions and thus of everyday life.

Societal Implications

Besides the theoretical implications of the current dissertation, the findings presented here also have some important practical implications. In Part I, I revealed that emphasizing moral implications of people's behavior caused them to inhibit their behavioral prejudice towards outgroup targets. This could imply that in real life settings, people may also adjust or control their behavioral or verbal expressions of prejudice when they are made aware of what such expressions might say about their own moral values. Consider for example a situation in which an employer rejects an applicant, merely because she indicates she wants to wear a headscarf at work. In this situation, the employer will probably only be aware of the consequences for the applicant rather than thinking about what the rejection of such applicants may reveal about himself and the company more generally. Awareness of the implications of his behavior in terms of his own morality and what it says about his values regarding equality and intercultural respect may make him more careful to ensure equal treatment in future interactions.

The findings reported in Part II of this dissertation also reveal that people care whether they are perceived as moral by *others*, especially by members of their own group. Motivating people to inhibit their prejudice towards outgroup targets by emphasizing moral values instead of competence, will thus be particularly effective within a group context. This finding speaks to debates about how to best promote diversity policies in work settings. In the literature, it has been suggested that, rather than emphasizing that diversity is the *right* thing to do, organizations should emphasize the 'business case for diversity'. In essence, this is a focus on *competence* that may persuade the executive board of a company to work towards a more diverse organization, and promotes ethnic and gender diversity in order to improve the organization's profit and success. In addition, the 'business case for diversity' is proposed to increase motivation and efficiency among employees (e.g., European Commission, 2005; Robinson & Dechant, 1997). However, in terms of diversity climate within the company, based on the results of this dissertation, I question the effectiveness of this measure in motivating employees to embrace diversity and treat colleagues from other ethnic backgrounds with respect. Instead, a better way to achieve this might be by emphasize that striving for a diverse organization is the *moral* thing to do.

Presenting policies in terms of moral principles, to motivate members to act accordingly may be a strategy that can actually be adopted by any kind of company, department, or team. Consider for instance organizations in the financial sector. Here, norms and performance targets also tend to be presented in terms of competence. In order to make a profit, close a successful deal, or attract new clients one should first and foremost be clever and skilled. Although this may sound intuitively convincing, my findings imply that it might be even more motivating for employees to be part of and work for an organization that emphasizes its moral character, for instance by focusing on fair treatment of employees, or showing honesty towards clients. Indeed, there is some correlational evidence in line with this reasoning, documenting that perceptions of organizational morality relate to employee satisfaction and work commitment (Ellemers, Kingma, Van de Burgt, & Barreto, 2011).

The notion that evaluations by other ingroup members are particularly effective in helping people to display moral behavior is also relevant in the context of the financial sector. In the Netherlands, organizations in this sector are being supervised and controlled by an external agency, the Authority of Financial Markets. This is an independent institution assigned to check and sanction the proper business conduct of financial markets, accountants or other financial service providers. However, an important consequence of this independent supervision is that evaluations of ethical business conduct are made by an *external* source. In the terminology used in my research, this would represent an outgroup judgment. That is, a judgment from a group that people tend to consider less self-relevant, which may for this reason alone be less effective in influencing their moral behavior. In view of the findings reported in this dissertation, it may be questioned whether supervision from such a source provides an optimal way to guide adherence to moral standards. If the goal is to improve morality in the financial sector, it might be more effective when moral norms are emphasized *within* a company and by its *own* board. Having moral business conduct as a core company value, is more likely to stimulate employees to perform their work in line with ethics guidelines.

Although emphasizing moral rather than competence norms may be particularly effective within one's own group, the findings of Part II of this dissertation also show that concerning people's control of prejudice, moral behavior can be influenced when this is evaluated an outgroup member. That is, people will generally be inclined to inhibit the expression of their negative bias when they are being monitored by a member of the group that is the target of such bias. This implies that diversity in a setting where people cooperate or evaluate one another may prevent displays of prejudice and discrimination. For example, having a Muslim employee as a member of an evaluation committee and who will thus observe the decision-making process concerning candidates for the job, may thus help the committee to create equal opportunities for Muslim as well as non-Muslim applicants. Likewise, having women present in the board of directors of a company may help others control gender bias when considering applications for high-status positions.

Importantly, my research also revealed that emphasizing the moral implications of people's behavior can be just as effective in reducing prejudice as is the presence of a representative of the group that is the target of prejudice. This is important for contexts in which intergroup contact is not feasible. This is the case for instance, when employees do not yet have any colleagues with a different ethnic background or religion, but might be induced in this way to be more open and welcoming to such a colleague. Likewise, emphasizing moral implications of being unbiased might be of benefit in the integration of newcomers in neighborhoods that primarily consist of people from the same social class or ethnicity. Within these contexts, people may be motivated to control their prejudice when this is emphasized as the right or moral thing to do, giving the new colleague or neighbor a fair chance to reveal their personal qualities rather than relying on biased expectations. Standard communications regarding company policy or national campaigns to encourage equal treatment of minority members tend to focus on the negative implications for the targets of prejudice, as a way to prevent people from expressing bias. My research suggests that there is likely to be added value in communicating about moral values and equality goals of the perpetrators, as a way to help diminish prejudice.

The research reported here not only elucidates how people adapt their moral behavior, it also reveals some very concrete and practical ways in which moral behavior can be stimulated. However, in real life, we have to take into account that even with the best of intentions people may sometimes deviate from what is considered moral, or be unable to always live up to their moral standards. The findings in Part III of this dissertation reveal how people are affected when confronted with their own moral slips. Because of the motivation to do what is morally right, confronting people with their moral failures has a negative impact upon their emotional well-being. If this negative response is sufficiently severe, it is likely to induce feelings of inadequacy and stress, which people are likely to cope with through denial or motivational withdrawal. Indeed, some of the fMRI evidence seems to suggest that negative moral information tends to be seen as less relevant to the self, even if skin conductance responses and self-reports indicate that receiving this type of information clearly has an emotional impact. Thus,

emphasizing their moral failures may not be the best way to motivate people to change or improve their behavior. Importantly however, people also seem to be especially attentive to positive information about their moral behavior which they seem to perceive as particularly relevant for their self-concept. This is relevant for instance to leaders who have to monitor and sanction the behavior of their subordinates. The natural tendency might be to confront an employee with moral failures, such as unethical decision making, as a way to prevent similar behavior in the future. However, due to the negative emotional impact this has, this might not be the best way to achieve behavioral change. Instead, it might be more effective to encourage the employee to succeed in their motivation to be moral by emphasizing moral achievements, or praising them for compliance with moral norms or company values while doing their job.

Thus far, I have mainly focused on the practical implications of the current findings in business settings. However, in principle, emphasizing the moral implications of people's performance can also be effective in stimulating moral behavior in other contexts. For example, similar mechanisms might be effective in sectors such as sports where moral behavior may be enhanced by emphasizing the importance of fair play and proper competition instead of focusing on winning outcomes alone.

Limitations and Directions for Future Research

The results of the empirical chapters outlined in Parts I, II, and III offer new insights in people's moral motivation. However, there are also some limitations in the studies described in Chapter 2-5 that need to be addressed and which may provide directions for future research.

A possible point of critique concerning the current dissertation is the repeated use of the implicit association test (IAT). In fact, this was the only (implicit) measure used to examine moral behavior in this thesis. In this research, the IAT was chosen because this measure lends itself rather well for framing its implications in terms of morality *and* in terms of competence. Also, the use of multiple trials makes it a measure of moral behavior that is also viable for the examination of cognitive responses, which requires repeated behavioral displays to achieve a reliable assessment of underlying processes. Moreover, the consistent use of the

IAT made it possible to compare and combine the related results of different scientific measures of cognitive processes and brain activation. Nevertheless, in future research other experimental paradigms might be developed to examine the importance of revealing one's motivation to be moral over one's motivation to be competent. For example, it would be interesting to examine whether people's own motivation to be moral rather than competent would also affect behavior in more economic type of situations, such as in bargaining games, where there is a clear trade-off between moral concerns (e.g., fairness, trust) and competence concerns (e.g., outcomes). That is, extending the current research as well as studies that examined the effects of knowledge about the moral character of the *other* player on the behavioral choices in such games (e.g., Delgado, Frank, & Phelps, 2005; Frank, Gilovich, & Regan, 1993), it could be tested whether morality is a stronger motivator than competence for people's own choices in situations where moral behavior may go at the expense of individual outcomes and ingroup norms trump individual gains.

Additionally, I have examined the importance of being perceived as moral by others, by introducing intra- and intergroup contexts. The examination of individual differences in the motivation to adhere to specific moral norms was not the focus of the current research. Nevertheless, previous research concerning prejudice control and automatic evaluative associations has revealed that such individual factors do explain differences in the regulation of social bias. For example, people can be internally and/or externally motivated to respond without prejudice on particular assessments (Plant & Devine, 1998; Amodio, Harmon-Jones, Devine, 2003; Amodio, Kubota, Harmon-Jones, & Devine, 2006). In some studies, participants are even preselected based on a measure of this motivation. For example, Amodio et al. (2006) recruited research participants who were previously found to score high on the internal motivation scale and low on the external motivation scale, to compare their responses with those of people who score high on both scales. In some of the chapters in this dissertation, an assessment of internal/external motivation to avoid prejudice was included as an additional background measure. In my research, participants generally showed internal instead of external motivation to appear unprejudiced. This is consistent

with the notion that I examined the motivation to be moral as a self-relevant goal. Future research might seek out research participants that are primarily externally motivated to appear unprejudiced. This might make it possible to examine for instance whether such individuals are less sensitive to feedback concerning their own morality, but might be more responsive to moral evaluations by others. Now that I have established these different concerns as relevant to the adaptation of moral behavior, it might be of interest to specify whether certain groups of individuals might be more open to certain types of moral interventions than others.

Conclusion

Using different scientific research methods, the findings in this dissertation reveal that (1) people tend to act in ways that are considered moral; (2) it is important for people to be perceived as moral by self-relevant others; and (3) that people care about succeeding in behaving according to their moral values. The findings extend previous research by observing and measuring people's actual behavior. Furthermore, automatic brain and physiological responses revealed how people respond to and initiate behavior in order to adhere to their moral values.

Appendices

Appendix A

Supplementary data Chapter 2

Pretest: Testing the Target Stimuli

Stimuli that represented the target concepts in our IAT consisted of 10 pictures of female faces without a headscarf and 10 pictures of female faces with a headscarf. All pictures were pretested by 67 participants (11 males), none of whom participated in the main study. Participants were asked to rate the pictures – that were presented as two groups: i.e., pictures of women with a headscarf and pictures of women without a headscarf were presented all on one screen – on personal characteristics, and ingroup (women without a headscarf) vs. outgroup (women with a headscarf) resemblance on a 7-point scale ranging from "not at all" to "to a great extent". Results showed that, although participants did not evaluate the women in the two groups differently concerning their perceived kindness, intelligence, competence, friendliness, genuinely, and trustworthiness, M(outgroup) = 5.00, SD = 0.63; M(ingroup) = 4.91, SD = 0.64; $t(66) = -1.33$, *ns*; they did report to perceive the women with headscarves to differ less from each other and to be more similar to each other than the women without headscarves; M(outgroup) = 3.74, SD = 1.51; M(ingroup) = 2.81, SD = 1.22; $t(66) = -5.41$, $p < .001$. Moreover, as intended, participants reported that they identified more with women without headscarves (the ingroup) than with women with headscarves (the outgroup); M(outgroup) = 2.60, SD = 1.03; M(ingroup) = 3.94, SD = 1.16; $t(66) = 7.96$, $p < .001$. The results thus indicated that, as intended, participants identified more with the ingroup. Furthermore, we found a clear ingroup/outgroup differentiation for women with and without a headscarf that is consistent with existing insights that outgroups tend to be perceived as more homogeneous than ingroups. This confirms that the stimuli we developed are suitable for our IAT.

A Pilot Study: Testing the IAT

Using two different task instructions, we framed the IAT as either a test of participant's morality or competence. However, although we argue that the IAT is an appropriate measure for the aim of our study, it is also possible that the test itself (without any additional information) raises morality concerns. After all, it could be evident for participants that a task concerning women with versus women without a headscarf has to do with prejudice or discrimination). We therefore first conducted a pilot study to test our new version of the IAT and to assess how the

test is interpreted by participants.

Method

Participants.

Twenty-six non-Muslim students from Leiden University (11 males, *M* age = 23.2 years, *SD* = 4.8) participated in the pilot study for money or course credits.

The implicit association test.

Stimuli. Besides the stimuli that represented the target concepts of the IAT (i.e., pictures of women with and without a headscarf; described in the pretest), there were also stimuli that represented the attributes. These consisted of 5 pictures of positive scenes (e.g., sun flowers), and 5 pictures of negative scenes (e.g., a tornado), selected from the International Affective Picture System (IAPS; Lang, Bradley, Cuthbert, 2005). The stimuli were selected based on the scores for pleasure (i.e., negative pictures with scores < 4 and positive pictures with scores > 7).

Experimental design.

The design of the IAT was identical to the design used by Greenwald, McGhee, and Schwartz (1998) in which the IAT consisted of 5 blocks. Congruent trials in test block 3 or 5 were trials for which female faces without a headscarf shared the same response key as positive pictures and female faces with a headscarf the same response key as negative pictures. Incongruent trials were trials for which female faces without a headscarf shared the same response key as negative pictures and female faces with a headscarf the same response key as positive pictures. The order of the congruent and incongruent trial blocks (3 and 5) was counterbalanced between participants. Blocks 1, 2, and 4 consisted of 26 trials and blocks 3 and 5 consisted of 156 trials each. Each trial started with a fixation point (with a duration that varied between 500-1500 ms), followed by stimulus presentation to which participants were supposed to respond (680 ms), and a feedback screen (500 ms). The feedback screen indicated whether participants responded correctly (indicated by a green check mark), incorrectly (i.e., a red cross), or whether they responded too late. Stimuli alternated between female faces and positive or negative pictures and the presentation order of stimuli was random. Participants could start each

block themselves and were thus able to take a short break in between. The experiment took approximately 25 minutes.

The IAT effect (*D* score).

The dependent measure was the IAT effect – indicated by the D score – calculated as the difference in reaction times on incongruent and congruent trials divided by a pooled standard deviation of all correct trials. This IAT effect was computed based on the scoring algorithm described by Greenwald, Nosek, and Banaji (2003). However, in contrast to IAT trials of Greenwald et al., where participants are asked to respond as quickly as possible but the stimuli only disappeared after a response was made, we used a limited presentation time of the stimuli (i.e., participants had to respond within 680ms after which the stimulus disappeared from the screen). We therefore did not have trials with extreme long or short latencies and we thus included them all, replaced error latencies with a replacement value (the mean plus two times the standard deviation of the correct latencies) and replaced zero latencies of the trials on which participants did not respond in time with the maximum response time of 680 ms.

Interpretation of the IAT.

After finishing the IAT we asked participants two questions (both positively and negatively formulated) concerning their interpretation of the IAT (i.e., "I think this test can assess my moral values concerning the equal treatment of different groups of people" / "I think this test cannot assess whether I am good in processing [new] information"). Participants could respond on a 7-point scale ranging from 1 "completely disagree" to 7 "completely agree".

Results and Discussion

Interpretation of the IAT.

Participants reported they were more inclined to think the test measured how well they are able to process new information (M = 4.27, SD = 1.34) than that the test measured their moral values concerning the equal treatment of different groups of people (M = 3.14, SD = 1.80); $t(25) = 3.44$, $p = .002$. This result thus negates our concern that the IAT raises morality concerns even though this is not made explicit.

IAT effect.

Participants showed the standard IAT effect (i.e., a negative implicit bias towards women with a headscarf); $t(25) = 2.61$, $p = .015$: Responding was more difficult on incongruent than on congruent trials (as was shown by increased reaction times and erroneous responses on incongruent compared to congruent trials). Our test thus revealed the typical IAT effect as it was first introduced by Greenwald et al. (1998).

The Instruction Manipulation

In the main manuscript, we shortly describe the difference between the two instruction conditions of our IAT. Here, we report the complete translation of these instructions.

Morality instruction.

"Is it important to you to treat people from different groups equally? Or do you have discriminating conceptions? Are you convinced that it is good to judge every individual, despite his or her gender, religion or ethnicity, in the same way? Or do you think it is right that some groups have a lower status in the Dutch society? People have different values concerning egalitarianism and discrimination. The test that you are about the do will show what kind of values you have and indicates whether your conceptions are discriminating against certain groups of people. The test is thus about important values you have and to what extent you strive for egalitarianism. The time to respond is limited, try to respond as quickly and as accurately as possible."

Competence instruction.

"Are you able to quickly and accurately respond to new information? Can you asses things very rapidly? Or, are you not able to quickly evaluate and respond to new information? People differ in how well they are able to pick up new information and how easy they can learn new tasks. The test that you are about to do will show how well you are able to process new information en indicates whether you can rapidly and accurately sort different types of pictures. This test is thus about sorting different types of images, a good performance and fast reaction times. The time to respond is limited, try to respond as quickly and as accurately as possible."

Appendix B

Supplementary data Chapter 4

Additional ERP results Study 4.2

Effects of Electrode Site

N1.

A main effect of electrode site for the N1 revealed that the N1 was greater at Cz (M = -7.44 μV, SE = 0.37) than at FCz (M = -6.66 μV, SE = 0.35); $F(1,56)$ = 14.84, $p < .001$, $\eta^2 = .21$. There was also a significant interaction between electrode, face and congruency; $F(1,56) = 3.92$, $p = .05$, $\eta^2 = .07$. Separate follow-up analyses revealed that there was a significant interaction between electrode and face on incongruent (and not on congruent) trials; $F(1,56) = 4.43$, $p = .04$, $\eta^2 = .07$, indicating that for incongruent trials the N1 modulation of viewing outgroup compared to ingroup faces was significant at Cz; $M_{difference}$ = -0.55, SE = 0.26, $F(1,56) = 4.34$, $p = .04$, $\eta^2 = .07$, but not at FCz; $M_{difference}$ = -0.07, SE = 0.21, $F < 1$.

P150.

The main effect of electrode site for the P150 showed that this ERP was greater at FCz (M = 5.13 μV, SE = 0.47) than at Cz (M = 4.08 μV, SE = 0.43); $F(1,56) = 75.65$, $p < .001$, $\eta^2 = .58$. There was also a significant interaction between electrode, face, congruency, task domain, and evaluator; $F(1,56) = 5.93$, $p = .02$, $\eta^2 = .10$. Follow-up analyses showed that (1) on incongruent (and not on congruent) trials there was an interaction between electrode, face, task domain, and evaluator; $F(1,56) = 7.04$, $p = .01$, $\eta^2 = .11$; (2) only at Cz (and not at FCz) there was a marginally significant interaction between face, task domain, and evaluator; $F(1,56) = 3.57$, $p = .06$, $\eta^2 = .06$. Separate analyses per task domain revealed a marginally significant face*evaluator interaction in the moral domain; $F(1,31) = 3.39$, $p = .08$, $\eta^2 = .10$, but not in the competence domain; $F<1$. Separate analyses per evaluator type revealed a marginally significant interaction between face and task domain in case of an outgroup evaluator; $F(1,27) = 3.14$, $p = .09$, $\eta^2 = .10$, but not in case of an ingroup evaluator; $F(1,27) = 1.02$, $p = .32$. Simple main effects revealed that the P150 modulation of enhanced social categorization was significant in the morality/ingroup condition ($F[1,31] = 12.84$, $p = .001$, $\eta^2 = .29$), but not in the morality/outgroup condition ($F < 1$). And significant in the competence/outgroup condition ($F[1,27] = 9.91$, $p = .004$, $\eta^2 = .27$), but not in the competence/ingroup condition ($F < 1$). Note that, besides the fact that we found these effects only at Cz

and incongruent trials, the increased P150 modulation in the morality/ingroup condition is consistent with our hypotheses and previous research (Van Nunspeet et al., 2014).

N450.

Results of the N450 also showed a main effect of electrode site; $F(1,56) = 86.49$, $p < .001$, $\eta^2 = .61$, indicating that the N450 was larger at CPz ($M = -0.44$ μV, $SE = 0.37$) than at Pz ($M = 0.95$ μV, $SE = 0.31$). There was also an interaction between electrode site and face; $F(1,56) = 22.05$, $p < .001$, $\eta^2 = .28$, indicating that the difference in the N450 between viewing non-Muslim (ingroup) compared to Muslim (outgroup) women was greater at Pz; $M_{difference} = -0.77$, $SE = 0.24$, $F(1,56) = 39.27$, $p < .001$, $\eta^2 = .41$, than at CPz; $M_{difference} = -1.38$, $SE = 0.22$, $F(1,56) = 22.05$, $p < .001$, $\eta^2 = .16$. Moreover, there was an interaction between electrode, congruency, and task domain; $F(1,56) = 4.42$, $p = .04$, $\eta^2 = .07$. However, follow-up analyses –separately for each electrode site and for each task domain condition– revealed no significant two-way interactions with congruency; F's < 2.29, p's > .14.

ERN.

For the ERN there was only a main effect of electrode site, revealing that the ERN modulation was greater at FCz ($M = -2.95$ μV, $SE = 0.47$) than at Cz ($M = -0.99$ μV, $SE = 0.46$); $F(1,44) = 76.20$, $p < .001$, $\eta^2 = .63$. There were no interaction effects with this factor.

The N450 Modulation of Viewing (non-)Muslim Faces

N450.

As the described in the main manuscript, we found a significant four-way interaction between congruency, face, domain and evaluator; $F(1,56) = 5.75$, $p = .02$, $\eta^2 = .09$. Since we were interested in the modulation of congruency, we included follow-up analyses examining this particular factor. However, we also found a main effect of faces: The N450 was larger for pictures of non-Muslim ($M = -0.28$ μV, $SE = 0.37$) compared to Muslim women ($M = 0.79$ μV, $SE = 0.33$); $F(1,56) = 24.06$, $p < .001$, $\eta^2 = .30$. We therefore also conducted analyses for the N450 modulations of faces: Separate analyses for the task domain conditions revealed a significant interaction between face, congruency, and evaluator in the morality condition; $F(1,31) = 5.36$, $p < .03$, $\eta^2 = .15$, but not in the competence

condition; $F(1,25) = 1.30$, $p = .27$. Furthermore, within the morality condition, there was an interaction between face and congruency in the ingroup evaluator condition; $F(1,16) = 10.26$, $p = .006$, $\eta^2 = .39$, but not in the outgroup evaluator condition; $F(1,15) < 1$. Simple main effects revealed that the N450 modulation of viewing non-Muslim compared to Muslim women in the morality/ingroup condition was significant on incongruent trials; $F(1,16) = 15.68$, $p = .001$, $\eta^2 = .50$, but not on congruent trials; $F < 1$.

References

Abraham, A. (2013). The world according to me: Personal relevance and the medial prefrontal cortex. *Frontiers in Human Neuroscience*, *7*, 341. doi:10.3389/fnhum.2013.00341

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, *68*, 804-825.

Amodio, D. M. (2010). Coordinated roles of motivation and perception in the regulation of intergroup responses: Frontal cortical asymmetry effects on the P2 event-related potential and behavior. *Journal of Cognitive Neuroscience, 22,* 2609-2617.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology, 94,* 60-74.

Amodio, D. M., Harmon-Jones, E., Devine, P. G., Curtin, J. J., Hartley, S. L., & Covert, A. E.(2004). Neural signals for the detection of unintentional race bias. *Psychological Science, 15,* 88-93.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*, 268-277. doi:10.1038/nrn1884

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323370. doi:10.1037//1089-2680.5.4.323

Beer, J. S., Stallen, M., Lombardo, M. V., Gonsalkorale, K., Cunningham, W. A., & Sherman, J.W. (2008). The Quadruple Process model approach to examining the neural underpinnings of prejudice. *NeuroImage*, *43*, 775-783.

Bengtsson, S. L., Lau, H. C., & Passingham, R. E. (2009). Motivation to do well enhances responses to errors and self-monitoring. *Cerebral Cortex*, *19*(4), 797-804. doi:10.1093/cercor/bhn127

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

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*, 624.

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering. *European Journal of Social Psychology, 41,* 135-143.

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2011). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology, 50,* 1-18.

Brett, M., Anton, J. L., Valabregue, R., & Poline, J. B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage*, *16*, 497.

Casebeer, W. D. (2003). Moral cognition and its neural constituents. *Nature Reviews Neuroscience*, *4*, 840-847.

Chee, M. W., Sriram, N., Soon, C. S., & Lee, K. M. (2000). Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. *Neuroreport*, *11*, 135-140.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*, 1249-1264. doi:10.1016/j.neubiorev.2012.02.008

Cope, L. M., Borg, J. S., Harenski, C. L., Sinnott-Armstrong, W., Lieberman, D., Nyalakanti, P. K., ... & Kiehl, K. A. (2010). Hemispheric asymmetries during processing of immoral stimuli. *Frontiers in Evolutionary Neuroscience*, *2*, 1-14. doi:10.3389/fnevo.2010.00110

Cocosco, C. A., Kollokian, V., Kwan, R. K. S., Pike, G. B., & Evans, A. C. (1997). Brainweb: Online interface to a 3D MRI simulated brain database. *NeuroImage*, 5, 425.

Crisp, R. J., & Hewstone, M. (1999). Differential evaluation of crossed category groups: Patterns, processes, and reducing intergroup bias. *Group Processes & Intergroup Relations*, *2*(4), 307-333. doi: 10.1177/1368430299024001

Crisp, R. J., & Hewstone, M. (2007). Multiple social categorization. *Advances in Experimental Social Psychology*, *39*, 163-254.

Crisp, R. J., Hewstone, M., & Rubin, M. (2001). Does multiple categorization reduce intergroup bias? *Personality and Social Psychology Bulletin*, *27*(1), 76-89. doi:10.1177/0146167201271007

Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. *Psychological Bulletin*, *87*(3), 546-563.

Cunningham, W. A., Van Bavel, J. J., Arbuckle, N. L., Packer, D. J., & Waggoner, A. S. (2012). Rapid social perception is flexible: Approach and avoidance motivational states shape P100 responses to other-race faces. *Frontiers in Human Neuroscience, 6,* 1-7. doi: 10.3389/fnhum.2012.00140

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

Dasgupta, N., DeSteno, D., Williams, L. A., & Hunsinger, M. (2009). Fanning the flames of prejudice: The influence of specific incidental emotions on implicit prejudice. *Emotion, 9,* 585-591.

Dawson, M. E., Schell, A. M., & Filion, D. L. (2000). The electrodermal system. In Cacioppo, J. T., Tassinary, L. G., & Berntson, G. G. (Eds.) *Handbook of Psychophysiology.*Cambridge University Press, Cambridge, 200-223.

Decety, J., & Porges, E. C. (2011). Imagining being the agent of actions that carry different moral consequences: an fMRI study. *Neuropsychologia*, *49*, 2994-3001. doi:10.1016/j.neuropsychologia.2011.06.024

Delgado, M. R., Frank, R. H., & Phelps, E. A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. *Nature Neuroscience*, *8*, 1611-1618. doi:10.1038/nn1575

Dickter, C. L., & Bartholow, B. D. (2007). Racial ingroup and outgroup attention biases revealed by event-related brain potentials. *Social Cognitive and Affective Neuroscience, 2,* 189-198.

Does, S., Derks, B., & Ellemers, N. (2011). Thou shalt not discriminate: How emphasizing moral ideals rather than obligations increases Whites' support for social equality. *Journal of Experimental Social Psychology, 47,* 562-571.

Does, S., Derks, B., Ellemers, N. & Scheepers, D. (2012). At the heart of egalitarianism: How morality framing shapes cardiovascular challenge versus threat in Whites. *Social Psychological and Personality Science, 3,* 747-753.

Dovidio, J. F., Kawakami, K., & Beach, K. R. (2001). Implicit and explicit attitudes: Examination of the relationship between measures of intergroup bias. In R. Brown & S. Gaertner (Eds.), *Blackwell handbook of social psychology: Intergroup processes* (pp. 175-197). Maiden, MA: Blackwell.

Ellemers, N., Kingma, L., Van de Burgt, J., & Barreto, M. (2011). Corporate Social Responsibility as a source of organizational morality, employee commitment and satisfaction. *Journal of Organizational Moral Psychology, 1*, 97-124.

Ellemers, N. & Van den Bos, K. (2012). Morality in groups: On the social-regulatory functions of right and wrong. *Social and Personality Psychology Compass, 6,* 878-889. doi: 10.1111/spc3.12001

Ellemers, N., Pagliaro, S., & Barreto, M. (2013). Morality and behavioural regulation in groups: A social identity approach. *European Review of Social Psychology*, *24*(1),160-193.

Ellemers, N., Pagliaro, S., Barreto, M., & Leach, C. W. (2008). Is it better to be moral than smart? The effects of morality and competence norms on the decision to work at group status improvement. *Journal of Personality and Social Psychology, 95,* 1397-1410.

European Commission (2005). The business case for diversity. Good practices in the workplace.http://ec.europa.eu/social/main.jsp?catId=370&featuresId=25&langId=nl.

FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012).What we say and what we do: the relationship between real and hypothetical moral choices. *Cognition*, *123*, 434-441. doi:10.1016/j.cognition.2012.02.001

Fiedler, K., & Bluemke, M. (2005). Faking the IAT: Aided and unaided response control on the implicit association tests. *Basic and Applied Social Psychology, 27,* 307-316.

Fiske, S. T., Cuddy, A. J. C., Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11,* 77-83.

Frantz, C. M., Cuddy, A. J. C., Burnett, M., ray, H., & Hart, A. (2004). A threat in the computer: The race implicit association test as a stereotype threat experience. *Personality and Social Psychology Bulletin, 30,* 1611-1624.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science, 4,* 385-390.

Gonsalkorale, K., Sherman, J. W., Allen, T. J., Klauer, K. C., & Amodio, D. M. (2011). Accounting for successful control of implicit racial bias: The roles of association activation, response monitoring, and overcoming bias. *Personality and Social Psychology Bulletin, 37,* 1534-1545.

Gratton, G., Coles, M.G., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography & Clinical Neurophysiology, 55,* 468–484.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 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,* 197-216.

Hajcak, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology, 42,* 151-160.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in cognitive sciences*, *4*, 223-233.

Hepper, E. G., Hart, C. M., Gregg, A. P., & Sedikides, C. (2011). Motivated expectations of positive feedback in social interactions. *The Journal of Social Psychology*,*151*, 455-477.

Ishai, A. (2008). Let's face it: it'sa cortical network. *Neuroimage*, *40*(2), 415-419. doi:10.1016/j.neuroimage.2007.10.040

Ito, T. A., & Urland, G. R. (2003). Race and gender on the brain: Electrocortical measures of attention to the race and gender of multiply categorizable individuals. *Journal of Personality and Social Psychology, 85,* 616-626.

Ito, T. A., & Urland, G. R. (2005) The influence of processing objectives on the perception of faces: An ERP study of race and gender perception. *Cognitive, Affective, & Behavioral Neuroscience, 5,* 21-36.

Jordan, A. H., & Monin, B. (2008). From sucker to Saint: Moralization in response to self-threat. *Psychological Science, 19*, 809-815. doi: 10.1111/j.1467-9280.2008.02161.x

Kawakami, K., Dovidio, J. F., Moll, J., Hermsen, S., & Russin, A. (2000). Just say no (to stereotyping): effects of training in the negation of stereotypic associations on stereotype activation. *Journal of Personality and Social Psychology*, *78*(5), 871-888. doi: I0.1037//O022-3514.78.5.871

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *The Journal of Neuroscience*, *32*, 16832-16844.doi:10.1523/JNEUROSCI.3016-12.2012

Kouzakova, M., Ellemers, N., Harinck, F., & Scheepers, D. (2012). The implications of value conflict: How disagreement on values affects self-involvement and perceived common ground. *Personality and Social Psychology Bulletin*, *38*, 798-807.

Kouzakova, M., Harinck, F., Ellemers, N., & Scheepers, D. (2014). At the heart of a conflict: Cardiovascular and self-regulation responses to value versus resource conflicts. *Social Psychological and Personality Science*, *5*, 35-42.

Kubota, J. T., Banaji, M. R., & Phelps, E. A. (2012). The neuroscience of race. *Nature Neuroscience*, *15*, 940-948. doi:10.1038/nn.3136

Kubota, J. T., & Ito, T. A. (2007). Multiple cues in social perception: The time course of processing race and facial expression. *Journal of Experimental Social Psychology, 43,* 738-752.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). International affective picture system (IAPS): Digitized photographs, instruction manual and affective ratings. *Technical Report A-6*. University of Florida, Gainesville, FL.

Lawrence, N. S., Wooderson, S., Mataix-Cols, D., David, R., Speckens, A., & Phillips, M. L. (2006). Decision making and set shifting impairments are associated with distinct symptom dimensions in obsessive-compulsive disorder. N*europsychology*, *20*, 409-419. doi:10.1037/0894-4105.20.4.409

Leach, C. W., Bilali, R., & Pagliaro, S. (2012). Groups and Morality. In J. Simpson & J. F. Dovidio (2013) (Eds.) *APA Handbook of Personality and Social Psychology, Vol. 2: Interpersonal Relationships and Group Processes.* Washington, DC: American Psychological Association.

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93,* 234-249.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of personality and social psychology*, *81*(5), 842-855. doi:1O.1O37//O022-35I4.8I.5.842

Lupfer, M. B., Weeks, M., & Dupuis, S. (2000). How pervasive is the negativity bias in judgments based on character appraisal?. *Personality and Social Psychology Bulletin*, *26*, 1353-1366. doi:10.1177/0146167200263004

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*, 1835-1838.

Martijn, C., Spears, R., Van der Pligt, J., & Jakobs, E. (1992). Negativity and positivity effects in person perception and inference: Ability versus morality. *European Journal of Social Psychology*, *22*, 453-463.

Monin, B., & Miller, D. T. (2001). Moral credentials and the expression of prejudice. *Journal of Personality and Social Psychology, 81,* 33–43.

Moran, J. M., Macrae, C. N., Heatherton, T. F., Wyland, C. L., & Kelley, W. M. (2006).Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience*, *18*, 1586-1594.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P. H., & Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology, 38,* 752-760.

Nigam, A., Hoffman, J. E., Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience, 4,* 15-22.

Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in cognitive sciences*, *8*, 102-107. doi:10.1016/j.tics.2004.01.004

Northoff, G., & Panksepp, J. (2008). The trans-species concept of self and the subcortical-cortical midline system. *Trends in Cognitive Sciences*, *12*, 259-264. doi:10.1016/j.tics.2008.04.007

Ochsner, K. N., Beer, J. S., Robertson, E. R., Cooper, J. C., Gabrieli, J. D., Kihsltrom, J. F., & D'Esposito, M. (2005). The neural correlates of direct and reflected self-knowledge. *Neuroimage*, *28*, 797-814. doi:10.1016/j.neuroimage.2005.06.069

Pagliaro, S., Ellemers, N., & Barreto, M. (2011). Sharing moral values: Anticipated ingroup respect as a determinant of adherence to morality-based (but not competence-based) group norms. *Personality and Social Psychology Bulletin, 37,* 1117-1129.

Pitcher, D., Walsh, V., & Duchaine, B. (2011). The role of the occipital face area in the cortical face perception network. *Experimental Brain Research*, *209*(4), 481-493. doi:10.1007/s00221-011-2579-1

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology, 75,* 811-832.

Poldrack, R. A. (2007). Region of interest analysis for fMRI. *Social Cognitive and Affective Neuroscience*, *2*, 67-70. doi:10.1093/scan/nsm006

Ratner, K. G., Kaul, C., & Van Bavel, J. J. (2013). Is race erased? Decoding race from patterns of neural activity when skin color is not diagnostic of group boundaries. *Social Cognitive and Affective Neuroscience, 8,* 750-755. doi:10.1093/scan/nss/063

Rebai, M., Bernard, C., & Lannou, J. (1997). The Stroop's test evokes a negative brain potential, the N400. *International Journal of Neuroscience, 91,* 85-94.

Reed II, A., & Aquino, K. F. (2003). Moral identity and the expanding circle of moral regard toward out-groups. *Journal of personality and social psychology*, *84*, 1270-1286. doi:10.1037/0022-3514.84.6.1270

Richeson, J. A., & Ambady, N. (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology*, *39*(2), 177-183. doi:10.1016/S0022-1031(02)00521-8

Robinson, G. & Dechant, K. (1997). Building a business case for diversity. *The Academy of Management Executive, 11,* 21 – 31.

Rudman, L. A., Dohn, M. C., & Fairchild, K. (2007). Implicit self-esteem compensation: automatic threat defense. *Journal of Personality and Social Psychology*, *93*(5), 798-813. doi:10.1037/0022-3514.93.5.798

Scheepers, D., Ellemers, N., & Derks, B. (2013). The "nature" of prejudice: What neuroscience has to offer to the study of intergroup relations. In: B. Derks, D. Scheepers, & N. Ellemers (Eds.). *The neuroscience of prejudice and intergroup relations*. New York: Psychology Press.

Schmitz, T. W., & Johnson, S. C. (2007). Relevance to self: A brief review and framework of neural systems underlying appraisal. *Neuroscience & Biobehavioral Reviews*, *31*, 585-596. doi:10.1016/j.neubiorev.2006.12.003

Skowronski, J. J., & Carlston, D. E. (1987). Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases. *Journal of Personality and Social Psychology, 52,* 689-699.

Stanley, D., Phelps, E., & Banaji, M. (2008). The neural basis of implicit attitudes. *Current Directions in Psychological Science*, *17*, 164-170. doi:10.1111/j.1467-8721.2008.00568.x

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American, 223,* 96-102.

Tajfel, H. (1978). *Differentiation between social groups*. London: Academic Press.

Urada, D., Stenstrom, D. M., & Miller, N. (2007). Crossed categorization beyond the two-group model. *Journal of Personality and Social Psychology*, *92*(4), 649-664. doi:10.1037/0022-3514.92.4.649

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: Evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience*, *23*(11), 3343-3354.

Van der Lee, R. (2013). Moral motivation within groups. Doctoral thesis. Leiden University.

Van der Meer, L., Costafreda, S., Aleman, A., & David, A. S. (2010). Self- and the brain: A theoretical review and meta-analysis of neuroimaging studies with implications for schizophrenia. *Neuroscience & Biobehavioral Reviews*, *34*, 935-946.doi:10.1016/j.neubiorev.2009.12.004

Van Nunspeet, F., Derks, B., Ellemers, N., & Nieuwenhuis, S. *(Manuscript under review).* Moral impression management: Evaluation by an ingroup member during a moral IAT enhances perceptual attention and conflict-monitoring.

Van Nunspeet, F., Ellemers, N., & Derks, B., *(Manuscript under review).* Reducing implicit prejudice against Muslim women: The effects of moral concerns, intra- and intergroup motives.

Van Nunspeet, F., Ellemers, N., Derks, B., & Nieuwenhuis, S. (2014). Moral concerns attention and response monitoring during IAT performance: ERP evidence. *Social, Cognitive, and Affective Neuroscience, 9,* 141-149. doi:10.1093/scan/nss118

Wegner, D. M. (2011). Setting free the bears: escape from thought suppression. *American Psychologist*, *66*, 671-680.

Williams, J. K., & Themanson, J. R. (2011). Neural correlates of the implicit association test: Evidence for semantic and emotional processing. *Social Cognitive and Affective Neuroscience, 6,* 468-476.

Willis, J., & Todorov, A. (2006). First impressions: making up your mind after 100-ms exposure to a face. *Psychological Science, 17,* 592-598.

Summary in Dutch

Samenvatting

Morele richtlijnen geven aan wat 'juist' en 'onjuist' gedrag is. Mensen vinden het belangrijk om moreel te zijn en moreel over te komen op anderen. Toch worden we allemaal wel eens geconfronteerd met mensen die immoreel gedrag vertonen; of doen we zelf wel eens iets waarvan we ons achteraf afvragen of dit wel het juiste was. In dit proefschrift heb ik onderzocht of, onder welke omstandigheden, en waarom mensen gemotiveerd zijn om zich moreel te gedragen. In voorgaand onderzoek werd voornamelijk het vermogen tot moreel redeneren en morele besluitvorming bestudeerd, om erachter te komen wat mensen *denken* dat een juiste handelswijze is. Ik bouw hierop voort, en richt mij op factoren die moreel *gedrag* stimuleren; wanneer, waarom, en hoe *doen* mensen wat ze juist vinden? Hierbij heb ik niet alleen het gedrag zelf onderzocht, of vertrouwd op wat mensen als redenen opgaven voor hun gedrag. Ik heb ook gekeken wat er gebeurt in het lichaam en in het hoofd van mensen die proberen moreel gedrag te vertonen. Hiermee kan ik meer informatie vergaren, zoals over gedachten en emoties waar mensen zelf geen zicht op hebben, of waarover ze mij niet willen vertellen. Ik heb hiervoor neurowetenschappelijke en psychofysiologische meetmethodes gebruikt om hersenactiviteit te meten (aan het schedeloppervlak met behulp van een elektrodencap, EEG; en in de hersenen met behulp van een MRI scanner; fMRI), en om te kijken of mensen het letterlijk 'warm' krijgen in bepaalde situaties (zweetreactie in huidgeleiding; SCR).

Dit proefschrift is opgebouwd in drie delen waarin telkens een andere vraag centraal staat. In Deel I heb ik bestudeerd of mensen geneigd zijn hun gedrag aan te passen of te controleren wanneer wordt benadrukt dat zij iets doen wat hun morele waarden kan onthullen. In Deel II heb ik bestudeerd of de motivatie om moreel gedrag te vertonen wordt beïnvloed, wanneer mensen worden geëvalueerd door anderen. In Deel III heb ik onderzocht of mensen het belangrijk vinden om zich te gedragen naar wat als moreel wordt beschouwd door te kijken hoe zij reageren op informatie die een indicatie geeft over de mate waarin hen dat is gelukt.

In Deel I heb ik onderzocht of mensen hun gedrag proberen aan te passen wanneer op voorhand wordt benadrukt dat de taak die ze doen iets zegt over hun morele waarden (in plaats van hun competentie). Dit heb ik onderzocht met behulp van een computertaak, een zogenaamde Impliciete Associatie Test (IAT), waarin

deelnemers worden gevraagd zo snel en accuraat mogelijk te reageren op verschillende soorten foto's en afbeeldingen. In mijn onderzoek liet ik foto's zien van vrouwen met en zonder hoofddoek, en van positieve en negatieve afbeeldingen (bijvoorbeeld een zonnebloem of tornado). De IAT is in eerder onderzoek gebruikt om onbewuste negatieve associaties bij bepaalde personen in kaart te brengen, die vooroordelen ten aanzien van sociale groepen kunnen onthullen. In de instructie voorafgaand aan de taak heb ik bij de helft van de deelnemers benadrukt dat hun prestatie iets kan zeggen over hun competenties (hoe goed zij zijn in het snel verwerken van informatie en het leren van nieuwe taken). Bij de andere helft van de deelnemers heb ik benadrukt dat hun gedrag aangeeft wat hun morele waarden zijn (wat betreft sociale gelijkheid en discriminatie). Resultaten in Hoofdstuk 2 laten zien dat deze instructie invloed heeft op het gedrag van mensen. Nadat is benadrukt dat hun prestaties op deze taak iets kunnen zeggen over hun moraliteit (in plaats van hun competentie), waren de onderzoeksdeelnemers meer geneigd zich moreel te gedragen. Dat wil zeggen: ze lieten minder negatieve vooroordelen zien ten aanzien van Moslima's bij het uitvoeren van deze taak.

Deze zelfde onderzoeksopzet heb ik herhaald, terwijl mensen een elektrodencap droegen (ERP studie – Hoofdstuk 2), of terwijl ze in de scanner lagen (fMRI studie – Hoofstuk 3). Hiermee kon ik hun hersenactiviteit meten tijdens het doen van deze taak. Event Related brain Potentials (ERP-maten) zijn hersengolven die laten zien *hoe sterk en hoe snel* mensen reageren op bepaalde gebeurtenissen, zoals de foto's die we ze laten zien, of de antwoorden die ze geven. Functionele hersenscans, gemaakt van het gehele brein terwijl mensen aan de taak werken (fMRI-maten) laten zien *welke delen* van de hersenen geactiveerd worden. Uit deze metingen van hersenactiviteit kunnen we dus afleiden waar mensen mee bezig waren tijdens de taak en welke cognitieve processen er (extra) worden geactiveerd om moreel gedrag te vertonen. Beide soorten metingen leveren dus ook aanvullende informatie over hoe en waarom mensen zorgen dat negatieve vooroordelen niet zichtbaar worden in hun gedrag, om te laten zien dat zij moreel zijn. Ik kan zo ook kijken of en hoe de hersenactiviteit tijdens de taak verandert, als mensen denken dat deze taak iets zegt over hun moraliteit, in plaats van hun competentie. De resultaten uit deze onderzoeken tonen aan dat mensen meer

aandacht hebben voor wie zij zien tijdens de taak (een Moslima of een vrouw zonder hoofddoek), als ze denken dat hun prestatie iets zegt over hun morele waarden. Dit wil zeggen dat ze meer geneigd zijn bij de foto's die in de taak getoond worden een onderscheid te maken tussen gezichten van vrouwen met of zonder hoofddoek (sociale categorisatie). De verhoogde neiging mensen te categoriseren in groepen, lijkt in eerste instantie misschien tegen-intuïtief als manier om vooroordelen tegen te gaan en gelijke behandeling te stimuleren. De taakprestaties suggereren echter dat de verhoogde aandacht voor het groepslidmaatschap van de vrouwen op de foto's de onderzoeksdeelnemers heeft geholpen om hun gedrag zodanig aan te passen dat zij beide groepen vrouwen gelijk konden behandelen. Ook laat de hersenactiviteit in de ERP resultaten van Hoofdstuk 2 zien dat mensen sterker reageren (wat betekent dat zij het erger vinden) als ze fouten maken tijdens de taak, wanneer zij denken dat hun prestatie aangeeft hoe moreel zij zijn. Tezamen tonen deze bevindingen dus aan dat mensen niet alleen *zeggen* dat zij het belangrijk vinden om moreel te zijn (zoals gebleken is uit eerder onderzoek), maar dat mensen ook *moeite doen* om zich daadwerkelijk te gedragen naar hun morele waarden.

In Deel 2 van dit proefschrift heb ik bestudeerd of de motivatie van mensen om moreel te zijn wordt beïnvloed door de aanwezigheid van anderen. In Hoofdstuk 4 laat ik zien dat het vooral van belang is om je moreel te gedragen in het bijzijn van iemand die tot jouw groep behoort. In dit gedeelte van het proefschrift dachten deelnemers dat zij geobserveerd werden tijdens het maken van de computertaak. Telkens nadat zij een respons gaven zagen zij een andere deelnemer die aangaf of zij de correcte of incorrecte respons hadden gegeven. Deze evaluator werd gepresenteerd als iemand met hetzelfde groepslidmaatschap als de deelnemer (iemand met hetzelfde persoonlijkheidstype, en dus een lid van dezelfde 'groep'), of als iemand met een ander groepslidmaatschap dan de deelnemer (iemand met ander persoonlijkheidstype en dus een lid van een andere groep). De resultaten van dit onderzoek lieten, net als in Deel 1, zien dat mensen minder negatieve vooroordelen ten aanzien van Moslims toonden wanneer de morele implicaties van hun gedrag waren benadrukt. Maar belangrijker was de nieuwe bevinding dat dit vooral gebeurde wanneer mensen tijdens hun prestatie op de computertaak werden geëvalueerd door iemand van hun eigen groep, en niet

wanneer zij werden geëvalueerd door iemand van een andere groep. Ook de hersenactiviteit die een rol speelt bij de motivatie om moreel gedrag te vertonen (en die ik eerder aantoonde in Hoofdstuk 2) was versterkt indien er een lid van de eigen groep meekeek tijdens de taak. Deze bevindingen laten dus zien dat moreel gedrag gestimuleerd kan worden door het benadrukken van de morele implicaties van dat gedrag, maar dat dit vooral effectief is wanneer er iemand meekijkt met wie we ons identificeren.

Het gedeelde of afwijkende groepslidmaatschap van de evaluator in Hoofdstuk 4 was gebaseerd op (fictieve) persoonlijkheidstypen. Dit groepslidmaatschap had geen betekenis buiten de onderzoeksruimte en was dan ook niet heel relevant. Dit geeft echter wel aan hoe sterk de motivatie is om moreel gevonden te worden door mensen die zijn zoals wij: Mensen vinden het zelfs belangrijk om moreel over te komen op iemand die zij niet kennen maar over wie hen enkel is verteld dat zij lid zijn van dezelfde groep (omdat ze dus iets met elkaar gemeen hebben zoals een persoonlijkheidstrek). In de computertaak draaide het echter om vooroordelen ten aanzien van Moslima's. De niet-Islamitische onderzoeksdeelnemers zullen zichzelf niet zien als lid van dezelfde groep als de Moslima's in de taak. Toch vroeg ik mij af of het mogelijk is dat zij zich wél moreler gaan gedragen wanneer hun taakprestatie wordt bekeken en beoordeeld door een vrouw met een hoofddoek. Dit heb ik dan ook onderzocht in Hoofdstuk 5. De resultaten lieten zien dat wanneer deelnemers werden geëvalueerd door een vrouw met een hoofddoek, zij minder negatieve vooroordelen ten aanzien van Moslims vertoonden. Sterker nog, wanneer deze evaluator was geïntroduceerd als iemand met hetzelfde persoonlijkheidstype als de deelnemer (en dus als een lid van dezelfde groep), konden deelnemers niet alleen hun negatieve associaties met Moslima's onderdrukken, maar ook hun positieve associaties met Moslima's versterken. De evaluatie door een vrouw met een hoofddoek was dus zeer effectief in het verminderen van negatieve vooroordelen over Moslims. Belangrijk is echter ook dat, zonder deze evaluator, negatieve vooroordelen verminderd werden als de implicaties van de prestatie van de deelnemers waren benadrukt in termen van moraliteit. Negatieve vooroordelen hebben vaak betrekking op minderheidsgroepen in de maatschappij. De kans is dus relatief klein dat iemands

gedrag door een lid van zo'n groep wordt beoordeeld. Het is dan dus van belang om te weten dat negatieve vooroordelen ook verminderd kunnen worden door het benadrukken van de morele implicaties van iemands gedrag. Tezamen tonen de resultaten van Deel 2 van dit proefschrift dus enkele manieren waarop negatieve vooroordelen ten aanzien van Moslims verminderd kunnen worden en hoe situationele factoren moreel gedrag kunnen beïnvloeden.

Nadat ik in Deel 2 van dit proefschrift had onderzocht of de aanwezigheid van anderen moreel gedrag kan beïnvloeden, keer ik in Deel 3 terug naar de persoonlijke motivatie van mensen om moreel te zijn. In Hoofdstuk 6 heb ik namelijk onderzocht of mensen het belangrijk vinden om te slagen in het vertonen van moreel gedrag. En of dit belangrijker is dan dat het hen lukt om zich competent te gedragen. In dit hoofdstuk heb ik mensen opnieuw de computertaak (de IAT) laten doen. Na afloop van de taak heb ik hen verteld dat de taak iets kan zeggen over hoe moreel en hoe competent zij zijn in vergelijking met anderen. Ook heb ik hen hun scores op de taak getoond. Terwijl de deelnemers hun scores zagen heb ik de huidgeleiding op hun handen gemeten om te testen of zij zich (onbewust) meer opwinden wanneer zij zien dat zij beter of slechter hebben gepresteerd dan andere mensen. De resultaten van deze metingen lieten zien dat mensen meer fysieke opwinding vertoonden wanneer zij te horen kregen dat zij minder moreel zijn dan anderen, dan wanneer zij vernamen dat zij minder competent zijn dan andere mensen. Ook gaven de deelnemers naderhand aan meer negatieve gevoelens te ervaren als zij hadden vernomen dat zij minder moreel zijn dan anderen. Als het mensen dus niet lukt om moreel gedrag te vertonen dan geeft dit hen een slecht gevoel.

In een vervolgstudie heb ik met behulp van fMRI onderzocht hoe de informatie over hoe moreel en competent mensen zich gedragen in vergelijking met anderen, verwerkt wordt in de hersenen. Dit keer heb ik mensen hun scores op de taak laten zien terwijl zij in de MRI scanner lagen. De resultaten toonden aan dat wanneer mensen hun testscores zagen, er een hersengebied werd geactiveerd waarmee we informatie detecteren die relevant is voor de vorming van ons zelfbeeld. Dit gebied werd tevens meer geactiveerd wanneer de deelnemers zagen dat zij moreler zijn in vergelijking met anderen dan wanneer zij zagen dat zij

competenter zijn dan anderen. De bevindingen in Deel 3 van het proefschrift tonen dus aan dat het nastreven van morele waarden en het vertonen van moreel gedrag belangrijk is voor hoe we onszelf zien. Het is pijnlijk om te moeten vernemen dat we minder moreel zijn dan anderen, maar vernemen dat we moreler zijn dan anderen is relevant voor de bepaling van ons zelfbeeld.

Conclusie

De bevindingen in dit proefschrift laten zien dat mensen het belangrijk vinden om zich te gedragen naar hun morele waarden. Zij vinden het belangrijk om moreel over te komen op anderen, vooral op mensen met wie zij zich kunnen identificeren zoals mensen die deel uitmaken van dezelfde groep. Daarnaast vinden mensen het belangrijk dat het hen lukt om zich moreel te gedragen en vinden zij het erger om te moeten constateren dat zij minder moreel zijn dan anderen, dan dat zij minder competent zijn dan anderen. Mensen hebben de motivatie om moreel te zijn. De kennis over hoe deze motivatie versterkt kan worden kan dan ook helpen het beste uit de mens naar boven te halen.

Acknowledgements

Dankwoord

Promoveren doe je gelukkig niet alleen. Ik heb van heel veel mensen mogen leren, en met vele anderen het bijbehorende plezier (of soms enige frustraties) kunnen delen. Er is echter meer waarvoor een woord van dank op zijn plaats is. Naomi: Dank voor je vertrouwen en alle kansen die je mij gegeven hebt. Je bent een geweldige begeleider, én een goede coach. Belle: Voor je kritische blik en de fijne samenwerking. Ik ben erg blij dat jij als copromotor bij dit project betrokken bent geweest. Sander: Voor de moed om aan de slag te gaan met nieuwe onderzoeksideeën en afwijkende designs. Evenals voor je enthousiasme en het delen van je kennis over ERPs en wetenschappelijk schrijven. Eveline: Voor jouw waardevolle bijdrage aan sommige hoofdstukken in dit proefschrift, maar ook voor je inspirerende werk waardoor mijn interesse in de neurowetenschap al snel werd gewekt. David: Thank you for your help with the development of new research ideas and questions, and sharing your knowledge about social neuroscience, in Leiden as well as in New York. Anna: Voor al je hulp bij onderzoek waarin jij gelukkig ook de uitdaging zag. Serge: Voor je expertise toen er zoveel vragen waren. Saïd en Welmer: Voor jullie support tijdens mijn verdediging, maar ook voor het kunnen delen van alle perikelen behorend bij het promoveren. De mensen van de technische ondersteuning, evenals collega's binnen de Cognitieve en de Ontwikkelingspsychologie (en later de Universiteit van Amsterdam): Dankzij jullie is er nooit een dag aan dataverzameling of -analyse verloren gegaan. Studenten, onderzoeksassistenten en 'confederates': Voor jullie hulp bij en het mogelijk maken van het verzamelen van de data. Collega's binnen de sectie S&O: Voor de gezelligheid en al jullie feedback tijdens meetings. Collega's van het Kurt Lewin Institute, en het Leiden Institute for Brain and Cognition: Voor alle leerzame cursussen en waardevolle discussies. All the contributors to the SISSA SCoNe Summer School in 2013 and the people from the Social Neuroscience Lab and the Social Perception and Evaluation Lab at NYU: Thank you for the great learning experience and the fun. Vrienden (zowel dank aan hen die ik heb opgedaan bij S&O, als aan hen die ik heb leren kennen tijdens mijn studietijd, of al op de middelbare school): Het is fijn om met jullie ook tijd te wijden aan andere zaken in het leven. Mijn grootouders: Voor jullie onvermoeibare interesse, zelfs al vertelde ik over zoiets als artefacten in neuroimaging data. Mijn zus en zwager: Zonder een

basis waarin ‘gezellig samen huiswerk maken aan de keukentafel’ en ‘hulp bij statistiek’ centraal stonden, had ik waarschijnlijk nooit gedaan wat ik nu doe. Mijn ouders: Omdat jullie er altijd zijn als ik jullie nodig heb, en anders ook. En Peter: Voor al je hulp bij de ‘Microsoft Office gestuurde’ aspecten van dit proefschrift. Maar bovenal voor je liefde en geduld; ik kijk ernaar uit nog meer van onze dromen samen uit te laten komen.

Curriculum Vitae

Félice van Nunspeet (The Hague, April 29th 1986) graduated from Hofstad Lyceum in The Hague in 2004. In 2007 she completed the Bachelor in Psychology and in 2009 the Research Master in Developmental Psychology, both at Leiden University. During her final year she investigated the neural correlates of social decision making in adolescents. In September 2009, Félice became a research assistant at the Social and Organizational Psychology Unit at Leiden University where she conducted pilot studies examining the neural underpinnings of moral motivation. These studies became part of her PhD research, which she started in June 2010 under supervision of Prof. Dr. Naomi Ellemers and Dr. Belle Derks. Félice is currently continuing her research as a post-doc at Leiden University, in the department of Social and Organizational Psychology.

Kurt Lewin Institute

Dissertation Series

The "Kurt Lewin Institute Dissertation Series" started in 1997. Since 2012 the following dissertations have been published:

2012-1: Roos Pals: *Zoo-ming in on restoration: Physical features and restorativeness of environments*

2012-2: Stephanie Welten: *Concerning Shame*

2012-3: Gerben Langendijk: *Power, Procedural Fairness & Prosocial Behavior*

2012-4: Janina Marguc: *Stepping Back While Staying Engaged: On the Cognitive Effects of Obstacles*

2012-5: Erik Bijleveld: *The unconscious and conscious foundations of human reward pursuit*

2012-6: Maarten Zaal: *Collective action: A regulatory focus perspective*

2012-7: Floor Kroese: *Tricky treats: How and when temptations boost self-control*

2012-8: Koen Dijkstra: *Intuition Versus Deliberation: the Role of Information Processing in Judgment and Decision Making*

2012-9: Marjette Slijkhuis*: A Structured Approach to Need for Structure at Work*

2012-10: Monica Blaga: *Performance attainment and intrinsic motivation: An achievement goal approach*

2012-11: Anita de Vries: *Specificity in Personality Measurement*\

2012-12: Bastiaan Rutjens: *Start making sense: Compensatory responses to control- and meaning threats*

2012-13: Marleen Gillebaart: *When people favor novelty over familiarity and how novelty affects creative processes*

2012-14: Marije de Goede: *Searching for a match: The formation of Person-Organization fit perceptions*

2012-15: Liga Klavina: *They steal our women: Outgroup Members as Romantic Rivals*

2012-16: Jessanne Mastop: *On postural reactions: Contextual effects on perceptions of and reactions to postures*

2012-17: Joep Hofhuis: *Dealing with Differences: Managing the Benefits and Threats of Cultural Diversity in the Workplace*

2012-18: Jessie de Witt Huberts: *License to Sin: A justification-based account of self-regulation failure*

2012-19: Yvette van Osch: *Show or hide your pride*

2012-20: Laura Dannenberg: *Fooling the feeling of doing: A goal perspective on illusions of agency*

2012-21: Marleen Redeker: *Around Leadership: Using the Leadership Circumplex to Study the Impact of Individual Characteristics on Perceptions of Leadership*

2013-1: Annemarie Hiemstra: *Fairness in Paper and Video Resume Screening*

2013-2: Gert-Jan Lelieveld: *Emotions in Negotiations: The Role of Communicated Anger and Disappointment*

2013-3 Saar Mollen: *Fitting in or Breaking Free? On Health Behavior, Social Norms and Conformity*

2013-4: Karin Menninga: *Exploring Learning Abstinence Theory: A new theoretical perspective on continued abstinence in smoking cessation*

2013-5: Jessie Koen: *Prepare and Pursue: Routes to suitable (re-)employment*

2013-6: Marieke Roskes: *Motivated creativity: A conservation of energy approach*

2013-7: Claire Marie Zedelius: *Investigating Consciousness in Reward Pursuit*

2013-8: Anouk van der Weiden: *When You Think You Know What You're Doing: Experiencing Self-Agency Over Intended and Unintended Outcomes*

2013-9: Gert Stulp: *Sex, Stature and Status: Natural Selection on Height in Contemporary Human Populations*

2013-10: Evert-Jan van Doorn: *Emotion Affords Social Influence: Responding to Others' Emotions In Context*

2013-11: Frank de Wit: *The paradox of intragroup conflict*

2013-12: Iris Schneider: *The dynamics of ambivalence: Cognitive, affective and physical consequences of evaluative conflict*

2013-13: Jana Niemann: *Feedback Is the Breakfast of Champions, but It Can Be Hard to Digest: A Psychological Perspective on Feedback Seeking and Receiving*

2013-14: Serena Does: *At the heart of egalitarianism: How morality framing shapes Whites' responses to social inequality*

2013-15: Romy van der Lee: *Moral Motivation Within Groups*

2013-16: Melvyn Hamstra: *Self-Regulation in a Social Environment*

2013-17: Chantal den Daas: *In the heat of the moment: The effect of impulsive and reflective states on sexual risk decisions*

2013-18: Kelly Cobey: *Female Physiology Meets Psychology: Menstrual Cycle and Contraceptive Pill Effects*

2013-19: Ellen van der Werff: *Growing environmental self-identity*

2013-20: Lise Jans: *Reconciling individuality with social solidarity: Forming social identity from the bottom up*

2013-21: Ruth van Veelen: *Integrating I and We: Cognitive Routes to Social Identification*

2013-22: Lottie Bullens: *Having second thoughts: consequences of decision reversibility*

2013-23: Daniel Sligte: *The functionality of creativity*

2014-01: Marijn Stok: *Eating by the Norm: The Influence of Social Norms on Young People's Eating Behavior*

2014-02: Michèlle Bal: *Making Sense of Injustice: Benign and Derogatory Reactions to Innocent Victims*

2014-03: Nicoletta Dimitrova: *Rethinking errors: How error-handling strategy affects our thoughts and others' thoughts about us*

2014-04: Namkje Koudenburg: *Conversational Flow: The Emergence and Regulation of Solidarity through social interaction*

2014-05: Thomas Sitser: *Predicting sales performance: Strengthening the personality – job performance linkage*

2014-06: Goda Perlaviciute: *Goal-driven evaluations of sustainable products*

2014-07: Saïd Shafa: In the eyes of others: *The role of honor concerns in explaining and preventing insult-elicited aggression*

2014-08: Félice van Nunspeet: *Neural correlates of the motivation to be moral*