Making sense of business failure: a social psychological perspective on financial and legal judgments in the context of insolvency

Strohmaier, N.

6 Are Morally Bad Directors More Culpable?

Effects of Directors' Moral Character on Mental State Ascriptions and Legal Judgments[1]

ABSTRACT

This chapter seeks to shed light on a phenomenon identified in experimental philosophy pertaining to people's tendency to judge actions as more intentional when these result in bad outcomes compared to good outcomes (i.e., the Knobe-effect, named after Joshua Knobe who first identified the phenomenon), and is therefore written with experimental philosophers in mind as the target audience. Nonetheless, considering the importance of judging directors' intentions in relation to acquiring new debt while approaching bankruptcy, this phenomenon has direct implications for how directors are judged. Despite being widely studied, the Knobe-effect remains poorly understood. In this chapter, I primarily aim to further our understanding of this effect in the context of directors' liability following a company's bankruptcy. Specifically, I seek to investigate to what extent moral character inferences play an important role in mental state ascriptions and legally relevant judgments such as blame, punishment, and foreseeability.

Summarizing this chapter in a more theoretically oriented manner, the starting point is that recent research in experimental philosophy has sought to understand how people conceive and use the legally important concept of intentional action. Despite significant advances over the past decade and a half, it remains puzzling why people judge actions that result in bad outcomes as more intentional than actions with good outcomes, and why different rules seem to apply for morally neutral events. In this chapter, I propose a novel account that can explain existing data while also providing new empirical evidence. In line with recent research in moral psychology supporting a personcentred approach to moral judgment, our Moral Character Account, as introduced in this chapter, suggests that mental state ascriptions as well as other judgments pertaining to concepts such as causality and freedom are ultimately driven by moral character inferences. Across five experiments (two preregistered; total $N = 1446$), we consistently found that morally bad directors are judged as acting more intentionally, knowingly, and recklessly than morally good directors in bringing about a harmful side-effect, and we also consistently

1 This chapter is based on: Strohmaier, N. & Kneer, M. Are Bad People More Culpable? Effects of Moral Character on Mental State Ascriptions in Legal Decision Making. In preparation.

found higher perceived likelihood of failure judgments as well as higher blame and punishment attributions for morally bad agents. Importantly, we found the same moral character effects among both lay people and legal professionals. Effects of outcome severity were limited and inconsistent, confirming the dominant role of moral character effects. Implications for theories of the folk psychology of intentional actions are discussed as well as the implications for legal practice.

6.1 Introduction

To determine whether someone deserves blame or praise, people have to assess that person's mental state as we generally assign less blame or praise for unintentional than for intentional acts. Assessing mental states accurately is paramount as it is undesirable to unduly blame people for harm they did not intent or even anticipate to elicit, nor do we want to give credit where it is not due. Mental state ascriptions are perhaps even more important in legal decision making. The Model Penal Code in the United States defines different levels of *mens rea* (guilty mind), and the degree of culpability is dependent on someone's specific mental state (i.e., purposely, knowingly, recklessly, negligently) at the time of the *actus reus* (guilty act). For example, intentionally harming someone is penalized harsher than when the harm is inflicted due to reckless or negligent behaviour. Differences exist across jurisdictions in determining culpability, but a universality is that legal sanctions depend to a relevant extent on a defendant's mental state when committing a transgression or crime. Considering the implications of mental state ascriptions both in daily life and in legal decision making, it is important that both lay people and legal professionals accurately assess other people's mental states.

Importantly, people have been shown to err when it comes to mental state ascriptions. In his seminal work, Knobe (2003a; Knobe, 2003b) demonstrated that people consider harmful behaviour as more intentional than helpful behaviour. Specifically, when a company director launched a program to boost the company's profits that, as a side effect, also either harmed or helped the environment, participants in his experiment believed more strongly that the director intentionally harmed the environment than that the director intentionally helped the environment. Such asymmetry in mental state ascriptions has been proven to be a robust phenomenon that replicates across a wide range of different cases (for comprehensive reviews, see Cova, 2016; Feltz, 2007), cultures (Knobe & Burra, 2006) and age groups (Leslie, Knobe, & Cohen, 2006), and also for mental states other than intentionality (e.g., Beebe & Buckwalter, 2010).

Asymmetries in mental state ascriptions have puzzled scholars ever since Knobe's initial experiment and a decade and a half later the phenomenon remains poorly understood. Nonetheless, many different explanations have

been put forward. A key distinction between the accounts that are currently out there is that some suggest that moral considerations (e.g., the badness of the outcome or blameworthiness of the agent) can explain the observed asymmetries (e.g., Cova, Lantian, & Boudesseul, 2016; Nadelhoffer, 2004b), whereas others argue that moral considerations do not play a role at all (e.g., Alfano, Beebe, & Robinson, 2012; Machery, 2008). Currently, the dominant view seems to be that moral considerations do in fact play an important role in intentionality judgments, but there is substantial disagreement regarding *how* exactly such considerations affect the folk psychology of intentional action (e.g., Cova, Lantian, & Boudesseul, 2016). For example, of the accounts incorporating moral considerations, there are those that consider such considerations to be indicative of the core concept of intentional action and therefore consider the asymmetry to be unproblematic (Knobe, 2006, 2010). In contrast, others have argued that any influence of moral considerations on intentionality ascriptions constitutes a bias (e.g., Alicke, 2008; Nadelhoffer, 2004a, 2004b, 2006). In short, the accounts put forward thus far come in many shapes and forms, yet there is currently no comprehensive theory that can fully explain the multitude of variations of the Knobe effect that exist today.

In this paper, we aim to achieve three goals. First, we aim to further our understanding of the Knobe effect and of the folk psychology of intentional action specifically and of mental state ascriptions more generally, by putting forward a theory that puts moral character evaluations at the centre of mental state ascriptions in morally laden contexts. We will argue that moral character evaluations bias mental state ascriptions such that these proportionally match the amount of blame an agent deserves based on his/her moral character. The account we put forward also attends to non-moral contexts in which moral character evaluations presumably play little or no role. In addition to putting forward our novel account of mental state ascriptions, the second goal of this paper is to provide empirical support for the biasing effect of moral character inferences in lay people's attributions of mental states. Finally, we aim to test whether legal professionals are just as biased as lay people in their mental state ascriptions following from moral character inferences. This would be particularly noteworthy as it substantiates the worries expressed about impartiality being at risk in legal cases (Nadelhoffer, 2006). Indeed, not only jury impartiality might be jeopardized due to the biasing effect of moral character evaluations, but also cases in which legal professionals have a more prominent role.

We first briefly describe the essence of our Moral Character Account (MCA), after which we use existing variations of the Knobe effect to explain our account in more detail.

6.1.1 Introducing the Moral Character Account of Intentional Action

Our account can probably best be introduced by describing the observed differences in intentionality ascriptions in two classic variations of the Knobe effect (Knobe, 2003b). The first variation concerns the following morally neutral case of a rifle contest in which the agent is either a skilled or unskilled marksman (text between brackets was varied between participants):

> Jake desperately wants to win the rifle contest. He knows that he will only win the contest if he hits the bull's-eye. He raises the rifle, gets the bull's-eye in the sights, and presses the trigger. *[Jake is an expert marksman. His hands are steady. The gun is aimed perfectly … The bullet lands directly on the bull's-eye. / But Jake is not very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild … Nonetheless, the bullet lands directly on the bull's-eye.]* Jake wins the contest.

Here, 79% of the participants presented with the skilled marksman version of the case said Jake intentionally hit the bull's-eye, whereas only 28% of those presented with the inexperienced marksman said the same. Hence, it appears that skill plays an important role in the folk concept of intentional action (see also Malle & Knobe, 1997). Now consider the following analogous case in which the agent kills a family member instead of hitting a target at a shooting range:

> Jake desperately wants to have more money. He knows that he will inherit a lot of money when his aunt dies. One day, he sees his aunt walking by the window. He raises his rifle, gets her in the sights, and presses the trigger. *[Jake is an expert marksman. His hands are steady. The gun is aimed perfectly … The bullet hits her directly in the heart. / But Jake is not very good at using his rifle. His hand slips on the barrel of the gun, and the shot goes wild … Nonetheless, the bullet hits her directly in the heart.]* She dies instantly.

In this morally laden case, the vast majority of the participants believed Jake intentionally killed his aunt, regardless of Jake's shooting skills (95% in the skilled version vs. 76% in the unskilled version). Hence, in morally laden cases, the perceived skill of the agent plays a much smaller role (if at all) in intentionality attributions.

To explain these differences in intentionality ratings between amoral and moral contexts, we follow other accounts adopting a multi-concept approach to the folk concept of intentional action (e.g., Cushman & Mele, 2008; Lanteri, 2010; Nichols & Ulatowski, 2007; Sousa & Holbrook, 2010) and suggest that people adhere to two distinct concepts. One concept is reserved for non-moral contexts, and the other for morally laden contexts. The former mirrors perceptions of *causal responsibility* and the latter mirrors those of *moral responsibility*. More specifically, in non-moral cases, when asked whether a certain event was brought about intentionally, we propose people ask themselves whether

the agent is causally responsible for bringing about the event. By causally responsible we mean whether there is a direct causal link between an agent's mental state and the event. For example, if you had the desire to pick up a cup standing right in front of you, merely wanting to pick it up is sufficient for the event (picking up the cup) to occur, hence why people will say you picked up the cup intentionally. Likewise, for a skilled marksman, merely wanting to hit the bull's-eye in a riffle contest is probably sufficient for hitting the bull's-eye, resulting in high intentionality ratings. In many cases, however, a certain desire is necessary but not sufficient to bring about an event. For the unskilled marksman in the riffle context case, a desire to hit the target was insufficient and Jake needed a large chunk of luck to hit the target, thereby weakening the causal link between his desire to hit the target and actually hitting the target, ultimately resulting in lower intentionality ratings. Our hypothesis is that the more additional causal factors (i.e., external to the agent) that are required for a certain event to occur, the weaker the perceived causal link between an agent's mental state (e.g., desire) and the effect and thus the lower the intentionally ratings will be.

In morally laden cases, we suggest the folk's concept of intentional action focusses on *moral responsibility*. The more morally responsible an agent is perceived to be for a certain event (good or bad), the more people will say the agent acted intentionally. Moreover, and this is crucial, we hypothesize that the key determining factor in assigning moral responsibility is people's evaluation of an agent's moral character. Once people have formed an opinion on the agent's character, they will want to blame morally bad agents in case of bad outcomes and be reluctant to give credit in case of good outcomes. This pattern will be reversed for morally good agents, such that people assign higher intentionality in case of good outcomes (such that the agent deserves praise), and lower intentionality in case of bad outcomes (as they are reluctant to blame the agent). Applied to the Jake killing his aunt case, we suggest participants considered Jake's moral character reprehensible given he is willing to kill his aunt for financial gains. People were then motivated to blame Jake and hold him morally responsible for the death of his aunt, not caring anymore about the strength of the causal link between Jake's wanting to kill his aunt and actually killing his aunt. Importantly, moral character perceptions need not be derived from an agent's motives for acting or the act itself, but can also stem from unrelated and irrelevant information. In short, to reach moral coherence, we suggest people will answer whatever question you give them in such a way that it is favourable for an agent with a good moral character and unfavourable for an agent with a bad moral character (Clark, Chen, & Ditto, 2015).

To summarize, we propose that the folk's use of the concept of intentional action depends on the context, such that people interpret intentionality in non-moral contexts in terms of an agent's causal responsibility for an outcome and for moral contexts in terms of moral responsibility, which is ultimately driven

by moral character evaluations. Figure 6.1 depicts a visual representation of our theory of intentional action. Before providing empirical support for the central role of moral character inferences in our theory, we first highlight previous studies that we believe can better be explained using our MCA and elaborate on the key differences and benefits relative to existing accounts.

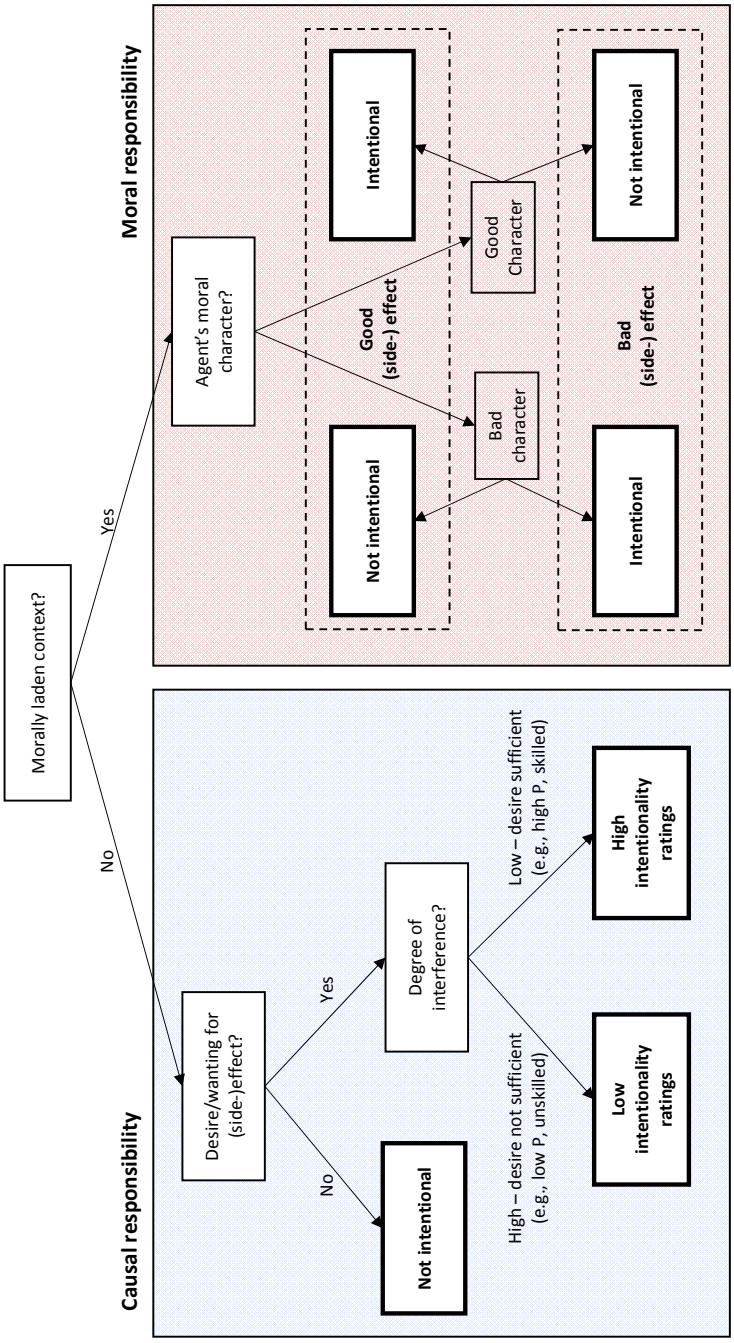

Figure 6.1. Dual concept theory of intentional action centring around causal and moral responsibility.

6.1.2 Using the MCA to explain existing data

In the original experiment conducted by Joshua Knobe, a salient feature is that the chairman looking to implement a program to boost his company's profits does not care at all about harming or helping the environment. This clearly speaks to the chairman's character and probably even more so for the chairman that does not care about harming the environment. Our MCA suggests that people will be pleased to say that the morally bad chairman intentionally harmed the environment as this implies the chairman deserves to be blamed and therefore gets what he deserves. In contrast, when participants in the help condition learned that the chairman helped the environment, they are reluctant to say the chairman intentionally helped the environment as this would imply the chairman deserves praise and they do not believe the chairman deserves praise at all. This is exactly what the original experiment showed. Hence, whereas Knobe originally explained the data in terms of the good or badness of the action, we suggest moral character evaluations drove the observed asymmetry in intentionality ascriptions.

A scenario that has been suggested to provide evidence for the idea that it *is* the inherent goodness or badness of an agent's behaviour (irrespective of blame) that drives the Knobe effect and not the blameworthiness of the agent is the following (Knobe & Mendlow, 2004):

> Susan is the president of a major computer corporation. One day, her assistant comes to her and says, "We are thinking of implementing a new program. If we actually do implement it, we will be increasing sales in Massachusetts but decreasing sales in New Jersey." Susan thinks, "According to my calculations, the losses we sustain in New Jersey should be a little bit smaller than the gains we make in Massachusetts. I guess the best course of action would be to approve the program." "All right," she says. "Let's implement the program. So we'll be increasing sales in Massachusetts and decreasing sales in New Jersey."

According to our MCA, people will interpret the question of intentionality in terms of the agent's causal responsibility for the side effect given that this particular case does not contain a moral element. The scenario is described in such a way that it seems that the president merely needs to give the go-ahead to her assistant to cause a decrease in sales in New Jersey, so no additional causal factors appear to be necessary for the side effect to occur. Hence, there is a strong causal link between the agent's mental state (i.e., desire to implement the program) and the side effect, which according to our theory would result in high intentionality ratings. This study indeed showed that 75% said the president intentionally decreased sales in New Jersey.

Wright and Bengson (2009) altered the above sales scenario in such a way that it more closely mimics the original chairman scenario, and as a result turned a morally neutral scenario into a morally laden scenario, which accord-

ing to our MCA therefore triggered judgments of moral responsibility and subsequent moral coherence processes. The scenario was as follows:

> The VP of a company went to the chairperson of the board and said, 'We are thinking of starting a new program. It will help us increase profits, *[but it will also decrease sales / and it will also increase sales]* in New Jersey.' The chairperson of the board answered, 'I don't care at all about *[decreasing / increasing]* sales in New Jersey. I just want to make as much profits as I can. Let's start the new program.' They started the new program. Sure enough, profits increased and sales in New Jersey *[decreased / increased]*.

The fact that the chairperson did not care about either decreasing or increasing sales does not speak favourably to the chairperson's character. It is safe to assume that a chairperson *should* care about such things and not caring in this case is a sign of bad corporate governance that might even provide ground for liability claims if shareholders were to suffer damages. Additionally, decreasing sales is generally considered to be a bad thing whereas increasing sales is generally considered to be a good thing. This can also be derived from the scenarios itself which stated: "but it will also decrease sales" versus "and it will also increase sales". It is therefore morally coherent to say the morally bad chairperson intentionally brought about the bad outcome (i.e., decreasing sales), but not to say the morally bad chairperson intentionally brought about the good outcome (i.e., increasing sales). The authors indeed found that intentionality ratings were higher in the *decrease* version of the scenario than in the *increase* version. In fact, the highest intentionality ratings were observed when respondents also considered the chairperson to be blameworthy for decreasing sales, which is in line with what our MCA would predict.

The original moral valance account (i.e., goodness or badness of the action drives the effect) has been critiqued by several other scholars who as an alternative mechanism suggested that norm violations are actually the driver of asymmetries in mental state ascriptions. They proposed that norm violations result in higher intentionality ratings than norm conforming actions (e.g., Alfano, Beebe, & Robinson, 2012; Hindriks, 2014; Holton, 2010; Robinson, Stey, & Alfano, 2015; Uttich & Lombrozo, 2010). However, these accounts cannot explain data showing that agents' attitudes towards an action or effect play a crucial role in intentionality judgments, regardless of the degree to which a norm is violated. Take for example Guglielmo and Malle's (2010) account which suggests that moral or normative considerations do not affect intentionality judgments at all and instead proposes that the Knobe effect can be fully explained by the agent's desire to achieve a certain outcome. They provided evidence for their account by for example showing that when a chairman regrets harming the environment, this significantly reduces intentionality ratings from 82% as observed in Knobe's original experiment to 40% in this regretful chairman scenario (see also Mele & Cushman, 2007; Phelan & Sarkissian, 2008).

To test whether an agent's desire for an outcome can indeed fully explain the Knobe effect or whether moral considerations also play a role, Cova et al. (2016) conducted several experiments in which they varied both the agent's attitudes towards the side effect (joyful vs. regretful) and the normative element of the side effect (harm vs. help). They showed that both the agent's attitude and the normative evaluation independently affected mental state ascriptions, thereby providing evidence for the notion that moral considerations do in fact impact intentionality judgments.

The observation that both moral considerations as well as agents' attitudes towards an outcome play a role in mental state ascriptions is wholly consistent with our MCA. In fact, we suggest that normative evaluations and agents' attitudes to be relevant for mental state ascriptions only in the sense that these factors are informative of the agent's moral character. Specifically, a chairman who does not care about or even derives pleasure from harming the environment can be considered to have a morally worse character than a chairman who regrets doing so, hence the lower intentionality ratings for the regretful chairman. Likewise, a chairman who is pleased to help the environment probably has a better moral character than a chairman who does not care at all about helping the environment, hence people's increased intentionality ratings for the morally good chairman when the environment was helped. Also, an agent who violates a norm will probably be judged as having a worse moral character than an agent conforming to that norm, especially when it concerns a salient norm (Robinson et al., 2015) that is important to the observer (Tannenbaum, Ditto, & Pizarro, 2007).

Importantly, accounts disregarding moral considerations and instead focusing on norm violations have in their defence put forward data showing intentionality judgments can be affected by non-moral norm violations. We suggest their data can in fact be explained by moral character evaluations. Take for example Knobe's (2004) vignette concerning an aesthetic norm violation, which reads as follows:

> The Vice-President of a movie studio was talking with the CEO. The Vice-President said: "We are thinking of implementing a new policy. If we implement the policy, it will definitely increase profits for our corporation, *[but it will also make our movies worse from an artistic standpoint / and it will also make our movies better from an artistic standpoint."]* The CEO said: "Look, I know that we'll be making the movies *[worse / better]* from an artistic standpoint, but I don't care one bit about that. All I care about is making as much profit as I can. Let's implement the new policy!" They implemented the policy. As expected, the policy made the movies *[worse / better]* from an artistic standpoint.

We consider it reasonable to assume that people will think a VP of a movie studio *should* care about a movie's artistic qualities and that people in general also do care about such qualities. Additionally, we consider it reasonable to assume that people tend to frown upon those who are so money-driven that

they have a complete disregard for relevant aspects of their work (i.e., the artistic qualities of a movie). Therefore, it is likely that people will evaluate the VP's moral character negatively. Assuming that making a movie artistically worse off is considered undesirable and improving a movie in an artistic sense is considered desirable, our theory would predict that when people evaluate the VP's moral character negatively, they will be reluctant to say the VP intentionally improved the movie and keen to say the VP intentionally made the movie worse. This is exactly what Knobe found. Only 18% of the participants said the VP intentionally improved the movie artistically, whereas 54% said the VP intentionally worsened the movie.

Another scenario concerning a norm violation that has previously been difficult to explain is the following (Knobe, 2007):

> In Nazi Germany, there was a law called the 'racial identification law.' The purpose of the law was to help identify people of certain races so that they could be rounded up and sent to concentration camps. Shortly after this law was passed, the CEO of a small corporation decided to make certain organizational changes. The Vice-President of the corporation said: 'By making those changes, you'll definitely be increasing our profits. *[But you'll also be violating the requirements of the racial identification law / But you'll also be fulfilling the requirements of the racial identification law].* The CEO said: 'Look, I know that I'll be *[violating / fulfilling]* the requirements of the law, but I don't care one bit about that. All I care about is making as much profit as I can. Let's make those organizational changes!' As soon as the CEO gave this order, the corporation began making the organizational changes.

The result showed that 81% of respondents said the CEO intentionally violated the requirements of the law whereas only 30% said he intentionally fulfilled the requirements of the law. We suggest that in the norm-violating condition, people consider the CEO to be morally good based on his blatant disregard for a norm-violating law. Since people will consider violating the racial identification law as good, it is morally coherent to say the CEO intentionally violated the law. In the norm-conforming condition, participants will probably be quick to judge the CEO as morally bad for not caring about complying with a norm-violating law. The side effect of fulfilling the requirements of the racial identification law is also considered bad, so on the surface it seems that moral coherence processes would result in participants saying the CEO intentionally complied with the norm-violating law. But that is not what Knobe found. Rather, only 30% said the CEO intentionally fulfilled the requirements of the racial identification law. An important element of our MCA is that the amount of moral responsibility assigned to the agent should be *proportionate* to the badness of the agent's character. Very bad agents deserve a lot of blame and mildly bad agents deserve a moderate amount of blame. In the scenario above, the CEO is clearly bad, but not as bad as a CEO who would endorse and be pleased to comply with such an extremely reprehensible law. The amount of blame implied by saying the CEO intentionally fulfilled the requirements of

the law would be disproportionate to the amount of blame the CEO deserved based on the evaluation of the CEO's character. Participants' thought processes might have been something along the following lines: "Look, this CEO surely is a bad person. He *should* care about whether or not he complies with this awful law. But I guess he is not a Nazi himself, because a Nazi *would* care about complying with that law and would even be happy to comply with the law and fully endorse the policy and principles behind it. That would be much worse. So even though I don't like this CEO very much, he did not really intentionally comply with the racial identification law as he is not that blameworthy of a person". Hence, based on the principle of proportionality, our theory is able to explain the asymmetry found in this particular scenario.

The notion of proportionality in our theory of intentional action is also supported by findings from Wible (2009) and Cova and Naar (2012). Wible found that intentionality ratings increased from 23% as found in the help condition of the original chairman case from Knobe (2003a) to 55% when the chairman who helped the environment as a side effect was actually a nice person and cared a lot about the environment. A further increase to 80% was found by Cova and Naar (2012) when the CEO was described as a *very* nice and altruistic person who was willing to help the environment without any financial or reputational gain.

To summarize, we have argued that the folk psychological concept of intentional action actually consists of two separate concepts (i.e., causal and moral responsibility) and that morally laden contexts trigger moral character evaluations which then drive mental state ascriptions. Furthermore, we have argued that proportionality is important in the sense that the blame implied by a certain mental state attribution should be proportionate to an agent's moral character. We have used existing data to argue for this central role of moral character evaluations and the element of proportionality. We will now briefly discuss the benefits of our account over alternative accounts and discuss how our account fits other recent research that has highlighted people's fundamental tendency to evaluate people's moral character.

6.1.3 Key benefits of the MCA

A prime benefit of our account is that it can explain data related to mental states other than intentionality. For example, Knobe (2004b) found similar asymmetries in his chairman case when he asked participants whether it was the chairman's *intention* to harm/help the environment. Likewise, using the same chairman case, Tannenbaum et al. (2007) found that participants were more inclined to say the chairman had a *desire* to harm the environment than to say the chairman had a desire to help the environment. Using analogous scenarios to the chairman case, Pettit and Knobe (2009) found similar effects for 'deciding', 'being in favour of', and 'advocating', with higher ratings for

the harm version than the help version. Beebe and Buckwalter (2010) also found asymmetries in the chairman scenario when they asked participants whether the chairman *knew* he would harm/help the environment, which they dubbed the 'epistemic side-effect effect' (see also Beebe & Jensen, 2012).

What all of these findings have in common is that the chairman (or other protagonists) in their scenarios clearly had a bad moral character. Each scenario also concerned a norm-violating or a norm-conforming side effect. Therefore, participants were probably keen to answer any question put to them in such a way that it signalled their discontent with the morally bad agent and such that it implied the appropriate amount of blame. After all, a morally bad chairman who desires to harm the environment is morally coherent, whereas a morally bad chairman with a desire to help the environment is not. The same goes for 'deciding', 'being in favour of', and 'advocating'.

Crucially, our account can also be extended to findings that were inspired by the Knobe effect but that do not concern mental states per se. Take for example the work of Knobe and Fraser (2008) on judgments of causality in which they presented participants with the following vignette:

> The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens. On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day, the receptionist needs to take an important message... but she has a problem. There are no pens left on her desk.

Participants were then asked to what extent they agreed with the following two statements: "Professor Smith caused the problem" and "The administrative assistant caused the problem". They found that participants were significantly more likely to agree with the former statement than the latter. Explaining these results in light of our MCA is relatively straightforward. The fact that the professor took a pen from the receptionist's desk while repeatedly being told not to signals a less than ideal moral character. It is morally coherent to then say the professor caused the receptionist's problem as this implies a certain degree of blame.

Phillips and Knobe (2009) also found asymmetries in judgments of an agent's freedom to perform a certain action. We argue that here too are moral character evaluations driving the effect. The scenario they used was the following:

> At a certain hospital, there were very specific rules about the procedures doctors had to follow. The rules said that doctors didn't necessarily have to take the advice of consulting physicians but that they did have to follow the orders of the chief

> of surgery. One day, the chief of surgery went to a doctor and said: "I don't care what you think about how this patient should be treated. I am ordering you to prescribe the drug Accuphine for her." The doctor *[had always disliked this patient and actually didn't want her to be cured / really liked the patient and wanted her to recover as quickly as possible].* However, the doctor knew that giving this patient Accuphine would result in *[an immediate recovery / her death]*. Nonetheless, the doctor went ahead and prescribed Accuphine. Just as the doctor knew she would, the patient *[recovered immediately / died shortly thereafter].*

When asked whether, given the rules of the hospital, the doctor did not really have the option of not prescribing Accuphine, participants were inclined to say the morally bad doctor who did not want the patient to be cured did not have the option of not prescribing Accuphine. This is morally coherent as the drug ultimately saved the patient (a good outcome), but participants probably did not feel the doctor deserved any credit for saving the patient as the doctor did not even want the patient to survive. By saying the doctor was forced to prescribe the drug, this precludes any credit for the doctor as he did not freely choose to prescribe the drug. In the scenario in which the doctor likes the patient and knows the drug will kill the patient but prescribes the drug anyway, participants probably felt this doctor has a weak moral character as the doctor should have stood up to the chief of surgery. It is then morally coherent to say the doctor *did* have the option of not prescribing the drug, as this implies the doctor deserves at least some blame for prescribing the fatal drug.

To conclude this point, we argue that our moral character account of the folk psychology of intentional action extends beyond 'mere' intentional action and can also explain related elements of folk psychology discussed above. As there is currently no theory or account that can explain all of the variations of the Knobe-effect, we consider our MCA to prevail over other theories that have been put forward, some of which have been addressed in this paper.

6.1.4 A Competing Theory

To further highlight benefits of the MCA over alternative accounts, we now compare the MCA with a competing theory put forward by Cova, Dupoux, and Jacob (2012). We focus on this account specifically as we believe it is, to date, the best candidate for explaining the Knobe-effect. Similar to our MCA, the authors take a pluralist stance and propose that people use different concepts of intentionally depending on the circumstances. In short, their first concept focusses on desire, such that agents are believed to act intentionally when their desire for a certain outcome is sufficiently strong. Their second concept suggests people will judge an agent to have acted intentionally when that agent is insufficiently reluctant to bring about an outcome than he/she is (normatively or descriptively) expected to be. Their third concept relates to skill and luck in such a way that people will judge an agent as acting

intentionally when the agent brings about an outcome by exerting control rather than by sheer luck.

We first identify some similarities between the two accounts. Their first and third concepts are similar to our causal responsibility approach to intentional action, in that we too propose that for an agent to act intentionally in a non-moral context, the agent should desire a certain event to occur and be able to exert sufficient control in bringing about the event (i.e., low degree of interference). In our MCA, however, both conditions need to be met for an agent to be judged as acting intentionally. Their second notion of intentional action shows some overlap with our account in that being less reluctant to perform a certain action than one might normatively expect speaks to a person's moral character. We therefore also consider reluctance in relation to normative expectations to be relevant, but only insofar as it can inform observers' evaluations of an agent's moral character.

Importantly, we consider our MCA to be more apt at explaining and predicting the folk psychology of intentional action for several reasons. First, whereas Cova et al. (2012) require three different concepts of intentional action to account for all the existent data, the MCA only requires two. We consider a more parsimonious account to be preferential if both achieve the same goal.

Second, whereas Cova et al. (2012, p. 390) state: "Arguably, there is no algorithm that would allow us to predict for each case which meaning will be the most salient", the MCA does in fact offer such an algorithm and a very basic one at that. Specifically, we argue that the concept of intentional action that people will use depends only on whether or not the situation at hand has a sufficiently salient moral element to it. We therefore believe that the MCA can better predict which concept of intentional action people will use.

Third, we question whether their account can actually account for all the existing data as they claim. For example, the authors introduce to their second concept of intentional action (i.e., being more or less reluctant than one might expect) the sensitivity to the value of a goal. In short, they argue that bringing about a bad side effect for a good reason suggests an agent is more reluctant to bring about the side effect than an agent bringing about the same bad side effect for mere futile reasons. In other words, they argue that motives for bringing about an effect are informative for assessing someone's reluctance towards bringing about that effect and that therefore motives are relevant for intentionality judgments (or at least for their second notion of intentionality that is based on reluctance). They use this line of reasoning to argue that their theory can account for the data presented by Nadelhoffer (2006), who describes two scenarios that are structurally identical but vary in terms of the agent's motive for bringing about a negative side effect. Specifically, in the first version, a thief knowingly endangers the life of a cop in an attempt to escape. Ultimately, the thief gets away and the cop dies as a result. In the second version, a man knowingly endangers the life of a thief in an attempt to escape from that thief who threatened the man with a gun. Ultimately, the man gets

away and the thief dies as a result. Cova et al. argue that the differences in intentionality ratings (37% said the thief intentionally brought about the cop's death versus 10% saying the man intentionally brought about the thief's death) can be explained by "the mere fact that protecting one's own good is a better goal than stealing the property of others".

Crucially, however, we question whether the goodness or badness of the goal in this case is informative with regards to the reluctance of the agents in bringing about the bad side effect, which according to the authors is the key element of their second concept of intentionality. In both scenarios it is explicitly stated that the agents do not care about endangering the cop's/thief's life and we consider it unlikely that the innocent man was more reluctant to endanger the thief who had just pulled a gun on him than the thief was reluctant to endanger the cop. Possibly, the man trying to get away from the thief in order to avoid being killed was actually less reluctant to risk the death of the thief than the thief was to risk bringing about the death of the cop merely to avoid jail time. Hence, based on Cova et al.'s account the data presented by Nadelhoffer is difficult to explain.

It is also unclear whether or not it matters if an agent is reluctant to bring about an adverse event for a normatively good reason (e.g., because endangering someone's life is bad) or for a mere pragmatic, selfish reason (e.g., limiting potential jail time). After all, someone can be *very* reluctant to bring about a bad event for selfish reasons. Our MCA has a clear prediction in this regard, as reluctance stemming from normatively good reasons speaks favourably to someone's character whereas reluctance stemming from selfish reasons does not. We therefore consider it more likely that moral character evaluations and subsequent moral coherence processes can account for Nadelhoffer's data. Specifically, the man trying to get away from the thief was presumably a good person or was judged as morally neutral considering the lack of information about the agent. Since bringing about someone's death is a very tragic event, people will be reluctant to say the morally good or morally neutral agent intentionally brought about the thief's death. In contrast, the thief was probably judged as having a bad moral character and intentionality ratings for bringing about the cop's death were therefore higher. Note however that only a minority said the thief intentionally brought about the cop's death. This can be explained by the MCA's element of proportionality. The thief 'only' stole some goods and his moral character was probably not judged as being so bad that it would warrant the blame implied by saying he intentionally brought about the cop's death. Our theory would predict that if the thief did not merely steal some goods but was rather a serial killer carrying a dead body in the trunk of his car, intentionality ratings for bringing about the cop's death would have been higher as this would then be more morally coherent.

A third limitation of Cova et al.'s account is that it presumably cannot explain findings related to mental states other than intentionality or concepts such as causality and freedom. Even though we believe the authors were solely

concerned with explaining intentionality judgments and therefore did not intend for their account to go beyond the scope of intentionality, as argued we consider it a strong point that our MCA cannot only explain the data related to intentionality, but also to other mental states and concepts as causality and freedom.

The final limitation of Cova et al.'s account that we identified is that even though in many cases it might be suitable for *predicting* intentionality judgments, we question whether it is also good at *explaining* these judgments. Specifically, we argue that our MCA is more likely to truly reflect folk psychological processes behind mental state ascriptions. Indeed, the literature has recently seen increased attention for moral character evaluations and its effects. For example, Pizarro and Tannenbaum (2012) recently argued that theories of moral judgment should include moral character evaluations as a key feature and they cite literature suggesting the motivation to evaluate others' moral character is a very primary and automatic psychological process that already manifests at a very early age and across cultures. Indeed, the notion that someone's moral character is the first aspect we determine when forming impressions of others has found empirical support (e.g., Goodwin, Piazza, & Rozin, 2014). There is also evidence that moral character evaluations even have primacy over perceptions related to a person's 'warmth' and 'competence' (Goodwin, 2015; Wojciszke, Bazinska, & Jaworski, 1998), which were previously believed to be the key elements in impression formation (e.g., Fiske et al., 2007).

Similar to Pizarro and Tannenbaum (2012) who argued that theories of moral judgments thus far are very *act-based* instead of *person-based*, so do we argue that theories of the folk psychology of intentional action have thus far largely neglected the central role of moral character evaluations. Considering that moral character evaluations are not incorporated in Cova et al.'s (2012) account of intentional action, we question whether it can accurately *explain* the folk psychology of intentional action rather than 'merely' *predict* intentionality judgments.

6.1.5 The MCA versus Existing Blame Accounts

Despite the lack of focus on moral character evaluations in theories of the folk psychology of intentional action, there have been several accounts that zoomed in on the blameworthiness of agents in mental state ascriptions. The most prominent one perhaps is that of Nadelhoffer, who has argued that the observed asymmetries in intentionality judgments can be explained by the fact that people (in the original chairman scenario) believe the chairman deserves blame for harming the environment and does not deserve praise for helping the environment, which then translates to intentionality judgments (Nadelhoffer, 2004b, 2004a). Likewise, Adams and Steadman (2004b, 2004a) have suggested that people merely use intentionality judgments as a way to

express their discontent with a blameworthy agent. Also, Alicke's blame validation assumption suggests that "once strong negative reactions have been evoked, people view the relevant evidence in a way that justifies their desire to blame the source of those reactions", including the degree to which an agent acted intentionally (Alicke, 2008; Alicke, 2000; Alicke, Weigold, & Rogers, 1990).

What these blame accounts of intentional action have in common, however, is that the blameworthiness of the agent is typically derived from his/her actions or motives for acting, such that certain acts, motives, or attitudes make an agent worthy of blame, rather than the agent's character per se. By such action-relevant motives or attitudes we mean for example the attitudes that the chairman had (not caring about the environment) regarding the side effect under consideration (harming or helping the environment). Crucially, our Moral Character Account goes one step further by suggesting that even information that is completely irrelevant to the action or (side-) effect under consideration, yet that is still informative to an agent's moral character, can affect mental state ascriptions.

Empirical evidence for the notion that moral character evaluations play an import role in blame attributions has been plentiful (Alicke & Zell, 2009; Critcher, Inbar, & Pizarro, 2012; Inbar, Pizarro, & Cushman, 2012; Nadler & McDonnell, 2012; Uhlmann & Zhu, 2014; Uhlmann, Zhu, & Diermeier, 2014), but the evidence for the effects of moral character inferences on intentionality judgments is scarce. Nonetheless, some studies do suggest that irrelevant information concerning an agent's character can affect intentionality ascriptions for an unrelated act. For example, Nadler (2012) showed that when an agent was described as having a bad moral character (e.g., unreliable, lazy, and unhelpful worker who arrives late for work or does not show up altogether), this agent was judged as having acted more intentionally (relative to an agent who was described as an exemplary employee and volunteer at an animal shelter) in relation to a fatal accident caused by the agent losing control while skiing and hitting the victim's head. Using different scenarios but a similar set-up, Nadler and McDonnell (2012) found further support for irrelevant character information affecting intentionality ascriptions. In addition to intentionality ascriptions, both these studies found that moral character inferences can also affect causal judgments, as well as judgments of blame, responsibility, and foreseeability.

Crucially, however, a few important questions remain. First, given the limited evidence for the effects of moral character inferences in intentionality judgments, it remains uncertain whether future studies will find similar effects. Second, and perhaps most importantly, despite the discussed importance of legal professionals getting mental state ascriptions right, there is as of yet no evidence whether legal professionals can also be affected by moral character evaluations in mental state ascriptions, especially when they are specialized in the subject matter that is used in a detail rich scenario. The latter would be the most stringent test of our MCA because legal professionals are presum-

ably aware of the fact that they should not let character information affect their judgments and because they are presumably also the least likely to rely on character information when they can instead rely on legally relevant information provided in the case, given their expertise on the matter. In this study, we aim to address both issues by further investigating the role of moral character evaluations in mental state ascriptions and other legally relevant judgments among both lay people and legal professionals.

Before introducing the experiments, we briefly turn to outcome effects in mental state ascriptions.

6.1.6 The MCA and Outcome Effects

In addition to the discussed factors that have previously been suggested to explain asymmetries in mental state ascriptions (e.g., norm violations, agents' attitudes, agents' blameworthiness), recent research has also investigated the role of outcome information. More specifically, a recent study by Kneer & Bourgeois-Gironde, 2017) tested whether in case of an adverse outcome, the severity of that outcome matters for intentionality ascriptions. They used a scenario in which a mayor is looking to start constructing a new highway connection, but who by doing so will either moderately harm the environment (animals in construction zone will be temporarily disturbed) or severely harm the environment (animals in the construction zone will die). Either way, the mayor does not care about harming the environment. The authors surveyed both lay people and professional judges and found that both groups were more inclined to say the mayor intentionally harmed the environment in case of the severe outcome relative to the moderate outcome.

An alternative explanation for these results (in light of our MCA) might be that participants considered the mayor who did not care about killing the animals to have a worse moral character than the mayor who did not care about temporarily disturbing the animals, and it might have been these moral character inferences which ultimately drove the observed effects in intentionality judgments. This leaves us with three potential hypotheses. First, it might indeed be that the observed severity effect was ultimately driven by moral character inferences. Second, it might be that severity effects constitute a unique effect distinct from that stemming from moral character inferences. Third, it might be that outcome severity and moral character inferences interact in some way when making mental state ascriptions. There is some evidence in support of the third hypothesis as a recent study found that the degree to which an action's consequences affect blame attributions is dependent on moral character evaluations, such that outcome information had a more pronounced effect on blame attributions for bad agents versus good agents (Siegel, Crockett, & Dolan, 2017). It remains uncertain, however, to what extent such findings can be translated to legally relevant judgments concerning for example mental

states, foreseeability, and punishment. Therefore, in this study we also investigate the potential relationship between outcome effects and moral character inferences in legally relevant judgments.

6.1.7 Current Study

To summarize, thus far we have presented our Moral Character Account of intentional action and (1) explained existing data in light of this account, (2) laid out its benefits relative to competing accounts, (3) explained how it differs from existing accounts incorporating agents' blameworthiness, and (4) hypothesized how it might be able to account for outcome effects in intentionality judgments. A few open questions remain that the experiments presented here seek to address. First, the evidence for intentionality judgments being affected by unrelated character information is scarce and non-existent for other mental states and other legally relevant judgments. Second, it remains unknown to what extent legal professionals will be affected by irrelevant moral character information when they are presented with a detail rich case in their own domain of expertise. Third, given the evidence for outcome effects among legal professionals in intentionality judgments, it remains unknown how such effects might relate to effects stemming from moral character information. In our studies we therefore aim to further test (1) the extent to which moral character inferences affect mental state ascriptions and other legally relevant judgments, (2) to what extent legal professionals are similarly affected as lay people in their judgments by character information, and (3) whether outcome severity and character information both affect lay people's and legal professionals' judgments, or whether these somehow interact.

In line with the original chairman scenario, we tested the above questions using a scenario concerning a director's liability for damages incurred by creditors following a decision made by the director. The legal background of this scenario is that when a company faces strong financial decline and possibly even insolvency, priorities should shift from creating shareholder value to protecting the company's creditors. When a company continues to do business and hence acquires further debt in the vicinity of insolvency, the company's directors may face personal liability for damages incurred by creditors when the company can ultimately not pay its debts. Conditions for being held liable as a director are that (1) at the time of incurring new debts, the company was as a certainty, or at least more likely than not, going to become insolvent, and that (2) the director knew or should have known that there was no reasonable prospect of paying for the debts. Under UK law, such a violation constitutes so-called 'wrongful trading'. Importantly, in most jurisdictions a distinction is made in some form or another between wrongful (or negligent) trading, which is mostly a civil offence, and fraudulent trading, which often constitutes a criminal offence (for an overview of the differences

across jurisdictions in the relevant laws pertaining to directors' liability, see INSOL International, 2017). In essence, the difference between wrongful trading and fraudulent trading is that a director can be held liable for fraudulent trading when he or she intentionally or recklessly incurs debts even though he or she knows there is no prospect of paying, whereas wrongful/negligent trading pertains to circumstances in which a director did not know but should have known that there was no reasonable chance of paying newly incurred debts. Hence, two key elements that have direct legal implications are the perceived likelihood at some point in time that a company will become insolvent, and the perceived mental state of the director at the time of incurring new debts.

The situation in which directors ought to shift priorities towards protecting their creditors is called the 'twilight zone' and frequently poses a moral dilemma. On the one hand, continuing to operate the business and trying to save it from insolvency risks liability if the turnaround attempt ultimately fails. On the other hand, informing third parties (e.g., creditors, new suppliers, lenders, etc.) of the company's dire financial situations risks creating a self-fulfilling prophecy (i.e., pushing the company into insolvency) as these third parties might immediately discontinue their operations with the financially distressed company for fear of not being paid.

In our studies, we presented both lay people (from the Netherlands and the US) and legal professionals (globally) with a case in which a director is faced with this dilemma and decides not to disclose any information concerning the company's financial state to its creditors; instead the director launches a final attempt to save the company from insolvency, which ultimately fails. To isolate the effects of moral character inferences from potential alternative explanations that have previously been suggested to explain mental state ascriptions, we kept constant (1) the degree to which the director deliberates his options and trade-offs he makes in the process (2) the director's perceived probability of his company going insolvent, (3) the fact that the director takes action to prevent the side effect from occurring (i.e., damages suffered by creditors), (4) the norm that was violated (i.e., protecting creditors), and (5) the director's motives for doing so (i.e., saving his own company). We varied the director's moral character (good vs. bad) as well as the severity of the outcome (moderate vs. severe). Study 1 and 2 were conducted among a sample of Dutch lay people and differed in terms of the director's subjective likelihood of failure (low vs. high). Study 3 aimed to replicate Studies 1 and 2 among a sample of US lay people after making several methodological improvements. Study 4 aimed to replicate Study 3 and provide more robust evidence for the effects of moral character inferences by adding legally relevant information to the case to further isolate moral character effects. Study 5 was identical to Study 4 and was conducted among legal professionals to allow for a direct comparison with lay people. This is particularly relevant for jurisdiction in

which legal professionals play a more prominent role in directors' liability investigations, such as is the case in for example the UK.

6.2 Study 1

6.2.1 Method

Participants, Design, and Procedure

Participants were recruited through a panel service located in the Netherlands. Based on an a-priori power analysis (using G*Power), we aimed to reach a final sample size of 210 to achieve a power of .95 to detect a medium effect size. In total, 326 participants completed the online survey as we oversampled to compensate for having to exclude participants from analyses. Participants who spent less than 60 seconds reading the case or who completed the survey within 120 seconds were excluded from the analyses. We used the same exclusion criteria across all studies. As a result, 72 participants were excluded, leaving a final sample of 254. The mean age was 49.6 (SD = 17.2) and 56.3% were female. None of the participants had a legal background.

We used a 2x2 between subjects design with moral character (bad vs. good) and outcome (moderate vs. severe) as factors. Participants were randomly assigned to one of the four conditions. The dependent variables were (1) three mental states (i.e., intentionality, knowledge, recklessness), (2) the perceived likelihood the company would ultimately go down and thus damage its suppliers, (3) three moral judgments (wrongness, blame, permissibility), and (4) the percentage of compensatory damages the director should pay (as a proxy for punishment).

Each participant was first presented with a business case (see Appendix 6.1 for the complete case) which described a company (suitcase manufacturer 'Yonos') facing financial difficulties and of which the director, named Devin, faced the dilemma between protecting his creditors versus trying to save his own company. The director was either described as having a good moral character (involved in setting up and running animal shelters) or a bad moral character (uses his influence to block the development of animal shelters). The director is informed by his advisor on the relevant laws concerning liability stemming from wrongful trading and that he is advised to inform the suppliers of the company's dire situation before placing any new orders. The advisor also asks how likely the director believes it to be the company will ultimately go down and damage the suppliers and the director answered he considers the chance of that happening to be very small. The director considers the risk of becoming insolvent as a direct result of disclosing the company's financial state to be bigger and decides to move forward with his plan to turn the company around. Next, the director instructs his advisor to place the necessary orders with the company's supplier and the director either does this reluctantly

as he has great relationships with his suppliers, or he does this without remorse as he does not care about his suppliers, consistent with the respective moral character conditions. Ultimately, the company becomes insolvent and the company's suppliers suffer damages as a result.

In the moderate outcome condition, these damages are limited and due to the suppliers' solid financial state they can easily take the hit. In the severe outcome condition, the suppliers could not take the hit due to their own financial issues and many went out of business as a result. Further, many employees were left without a job. It was stated explicitly that neither the director nor the advisor could oversee how badly the suppliers might be affected. In an attempt to answer the call for more detail rich materials in studying mental state ascriptions (e.g., Alicke, 2008; Young & Cushman, 2006), this case was relatively elaborate and contained details regarding the company's background, its financial status, the director's turnaround plan, etc. The complete case including the character and outcome manipulations can be found in Appendix 6.1.

Participants were then asked to indicate on Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree) to what extent they agreed with the following statements: (1) "The director intentionally damaged the suppliers", (2) "When the director placed the orders, he knew practically certain he would damage the suppliers", (3) "When the director placed the orders, he knew there was a substantial risk he would damage the suppliers", (4) "When the director placed the orders, he did not know but should have known there was a substantial risk he would damage the suppliers". These questions were presented on separate screens and in a random order and corresponded to the mental states purposefully, knowingly, recklessly, and negligently, respectively, as described in the Model Penal Code. Next, participants were asked to answer the degree to which they believed the director's decision to not inform his suppliers before placing the orders to be wrong, blameworthy, and impermissible. These three moral judgments were presented in random order and were answered on a 7-point scale, ranging from (1) not at all wrong/blameworthy/impermissible, to (7) very wrong/blameworthy/impermissible. Finally, participants were asked to indicate on a scale from 0-100% what percentage of the total damages incurred by the suppliers should be paid by Devin (as a measure of punishment) and how likely they believed it to be (also on scale from 0-100%) that the company would ultimately fail and not be able to pay for the orders they placed with the suppliers.

To check whether the moral character manipulation worked, participants were asked using a 7-point scale whether they considered the director's moral character to be (1) "very good" or (7) "very bad". Likewise, to check whether participants indeed considered the severe outcome to be worse than the moderate outcome, participants were at the end of the survey presented with each scenario and asked to indicate on a 7-point scale whether they considered the outcome to be "not bad at all" (1) or "very bad" (7)

6.2.2 Results

The manipulation check showed that participants in the bad moral character condition rated the director as morally worse than participants in the good moral character condition ($M = 6.27$, $SD = 1.02$ vs. $M = 4.42$ $SD = 1.32$), $F(1,250) = 146.63$, $p < .001$, $\eta_p^2 = .370$. Outcome severity did not affect perceptions of moral character, $F(1,250) = 2.11$, $p = .147$, $\eta_p^2 = .008$. The manipulation check for outcome severity showed that participants considered the severe outcome to be worse than the moderate outcome ($M = 5.75$, $SD = 1.78$ vs. $M = 3.94$, $SD = 1.60$), $t(253) = 13.99$, $p < .001$, $d = 0.88$.

Next, a Multivariate Analysis of Variance (MANOVA) was conducted with the variables pertaining to the four mental states as dependent variables and moral character and outcome severity as independent variables. A significant effect was found for moral character, $F(4,247) = 6.95$, $p < .001$, $\eta_p^2 = .101$, but not for outcome severity, $F(4,247) = 0.67$, $p = .616$, $\eta_p^2 = .011$, nor for the interaction effect, $F(4,247) = 1.23$, $p = .298$, $\eta_p^2 = .020$. Subsequent Univariate Analyses of Variance (ANOVA) showed that moral character had an effect on all mental states, apart from negligence. Specifically, participants judged (1) the bad director to have acted more intentionally in damaging the suppliers than the good director ($M = 5.09$, $SD = 1.65$ vs. $M = 4.14$ $SD = 1.64$), $F(1,250) = 18.61$, $p < .001$, $\eta_p^2 = .069$, (2) the bad director to have knowingly caused the damages of the suppliers to a larger extent than the good director ($M = 4.77$, $SD = 1.64$ vs. $M = 3.95$, $SD = 1.55$), $F(1,250) = 15.11$, $p < .001$, $\eta_p^2 = .057$, and (3) the bad director to have acted more recklessly in damaging the suppliers than the good director ($M = 6.11$, $SD = 0.93$ vs. $M = 5.58$, $SD = 1.22$), $F(1,250) = 12.36$, $p = .001$, $\eta_p^2 = .047$). No difference between the bad and good director was found for negligence ($M = 4.61$, $SD = 2.22$ vs. $M = 4.72$, $SD = 1.78$), $F(1,250) = 0.49$, $p = .487$, $\eta_p^2 = .002$. Figure 6.2 provides a visual presentation of the findings.

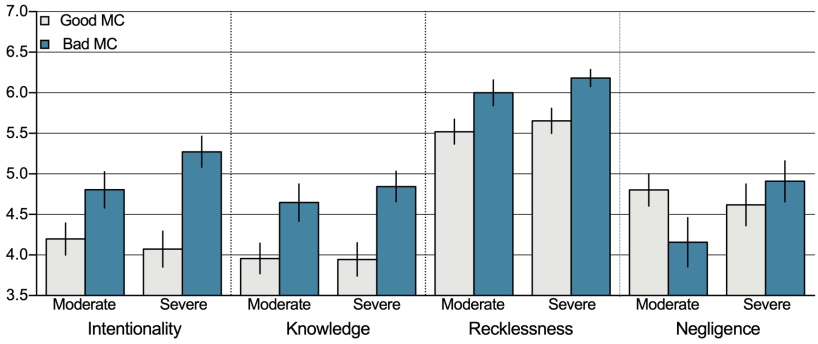

Figure 6.2. Mean scores for the four mental states, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

Regarding the perceived likelihood that the company would ultimately fail and damage its suppliers, an effect was found for moral character $F(1,250) = 9.86$, $p = .002$, $\eta_p^2 = .038$, but not for outcome severity, $F(1,250) = 0.48$, $p = .489$, $\eta_p^2 = .002$, nor for the interaction between the two factors, $F(1,250) = 0.76$, $p = .383$, $\eta_p^2 = .003$. Participants considered the likelihood of failure to be higher in the case with the morally bad director compared to the morally good director ($M = 68.30$, $SD = 24.05$ vs. $M = 58.75$, $SD = 22.44$). Similarly, only an effect of moral character was found for punishment (percentage of damages to be paid by the director), such that participants assigned a higher percentage in the case with the morally bad director than with the morally good director ($M = 79.50$, $SD = 22.61$ vs. $M = 69.53$, $SD = 24.69$), $F(1,250) = 9.92$, $p = .002$, $\eta_p^2 = .038$. No effect was found for outcome severity, $F(1,250) = 2.05$, $p = .152$, $\eta_p^2 = .008$, nor for the interaction, $F(1,250) = 2.53$, $p = .113$, $\eta_p^2 = .010$. The findings for the likelihood of failure and punishment are presented in Figure 6.3.

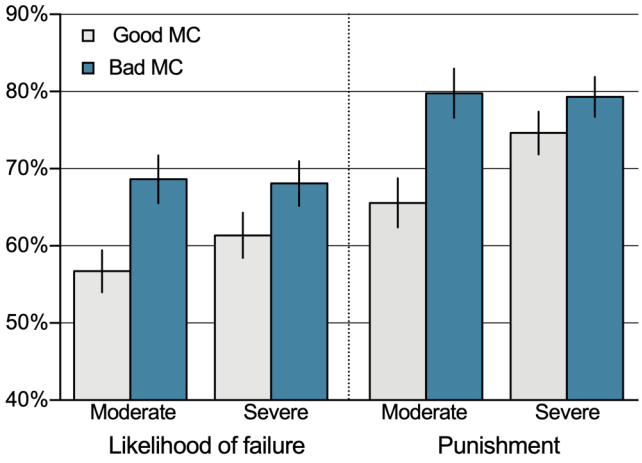

Figure 6.3. Mean scores for the perceived likelihood of failure and punishment, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

In relation to the moral judgments, a MANOVA with all three measures as dependent variables returned a significant effect for moral character, $F(3,248) = 8.34$, $p < .001$, $\eta_p^2 = .092$, and for outcome severity, $F(3,248) = 3.02$, $p = .030$, $\eta_p^2 = .035$, but not for the interaction, $F(3,248) = 0.44$, $p = .728$, $\eta_p^2 = .005$. The univariate analyses showed that the director's actions were judged as more wrong ($M = 6.16$, $SD = 0.97$ vs. $M = 5.47$, $SD = 1.23$), $F(1,250) = 20.88$, $p < .001$, $\eta_p^2 = .077$, more blameworthy ($M = 6.08$, $SD = 1.14$ vs. $M = 5.56$, $SD = 1.17$), $F(1,250) = 10.13$, $p = .002$, $\eta_p^2 = .039$, and more impermissible ($M = 5.94$, $SD = 1.37$ vs. $M = 5.11$, $SD = 1.38$), $F(1,250) = 20.40$, $p < .001$, $\eta_p^2 = .075$, when the case involved a morally bad director compared to a morally good director.

Regarding outcome severity, the director's actions were judged as more wrong (M = 6.03, SD = 0.98 vs. M = 5.58, SD = 1.29), $F(1,250) = 6.06$, $p = .014$, $\eta_p^2 = .024$, and more blameworthy (M = 6.05, SD = 1.03 vs. M = 5.57, SD = 1.29), $F(1,250) = 7.41$, $p = .007$, $\eta_p^2 = .029$ when the case had a severely bad outcome compared to a moderately bad outcome. No effect of outcome severity was found for the permissibility of the director's actions, $F(1,250) = 2.11$, $p = .147$, $\eta_p^2 = .008$. The findings for moral judgments are presented in Figure 6.4.

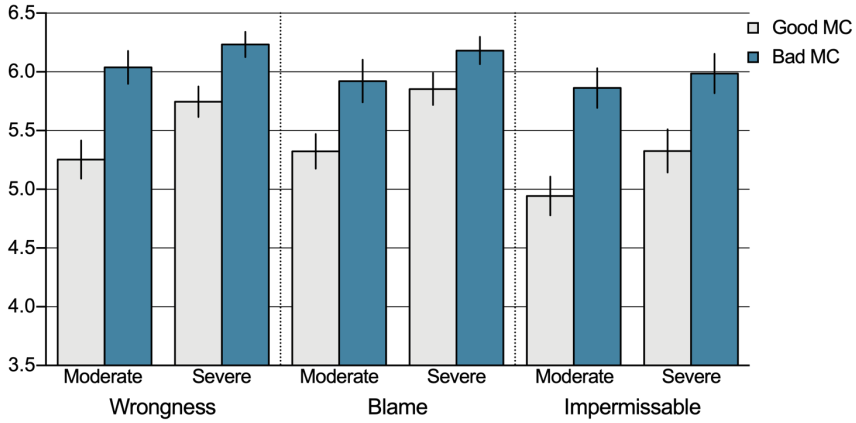

Figure 6.4. Mean scores for the three moral judgments, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

6.2.3 Discussion

Study 1 showed that moral character inferences have a significant influence on mental state ascriptions, as well as the perceived likelihood of failure, punishment, and moral judgments. Only for negligence did moral character not have any effect. This might be due to the fact that people generally considered the director to have acted recklessly. When a participant strongly agrees with the statement "the director knew there was a substantial risk...", it would be odd for this person to then also agree with "the director did not know but should have known there was a substantial risk …". More likely would be to find a reversed effect for negligence, as high attributions of intent, knowledge and recklessness should be accompanied with low attributions of negligence. Interestingly, contrary to previous research that found an effect for outcome severity on intentionality ascriptions (e.g., Kneer & Bourgeois-Gironde, 2017) in the present study outcome severity only affected attributions of blame and wrongness judgments. It seems therefore that when specific information regarding an agent's moral character is provided, outcome severity has very little additive effect.

6.3 STUDY 2

Study 2 aimed to replicate the findings of Study 1 and also to investigate the role of subjective probability. That is, whereas in Study 1 the case stated that the director believed the chance of failure (and thus damaging the company's suppliers) was very small, in Study 2 the director believes the chance of failure is actually very large. Study 2 therefore provides a more stringent test of the effects of moral character inferences as in this study it is explicitly stated the director knows there is a substantial risk of failure and the director even believes the chance this will materialize is very large. Apart from the director's subjective probability of failure, Study 2 is identical to Study 1. However, as the director believes his chance of failure is very large, the dependent variable measuring negligence (i.e., "the director did not know but should have known there was a substantial risk…" becomes obsolete; it was therefore omitted in this study.

In total, 339 participants completed the online survey as we oversampled to compensate for having to exclude participants from analyses (based on the same criteria as in Study 1). Of the original sample, 100 participants were excluded leaving a final sample of 239. The mean age was 49.7 (SD = 17.2) and 53.6% were female. Participants were again recruited using the same Dutch online panel service. None of the participants in Study 2 took part in Study 1 and none had a legal background.

6.3.1 Results

The manipulation check again showed that participants in the bad moral character condition rated the director as morally worse than participants in the good moral character condition (M = 6.22, SD = 1.06 vs. M = 4.50 SD = 1.30), $F(1,235) = 123.96$, $p < .001$, $\eta_p^2 = .345$. Outcome severity again did not affect perceptions of moral character, $F(1,235) = 1.97$, $p = .162$, $\eta_p^2 = .008$.

Next, a MANOVA with the variables measuring the three mental states as dependent variables and moral character condition and outcome severity as independent variables revealed a significant effect for moral character, $F(3,233) = 12.04$, $p < .001$, $\eta_p^2 = .134$, but not for outcome severity, $F(3,233) = 0.45$, $p = .717$, $\eta_p^2 = .006$, nor for the interaction effect, $F(3,233) = 1.52$, $p = .209$, $\eta_p^2 = .019$. Subsequent ANOVAs showed that participants judged (1) the bad director to have acted more intentionally in damaging the suppliers than the good director (M = 5.30, SD = 1.43 vs. M = 4.52 SD = 1.49), $F(1,235) = 16.70$, $p < .001$, $\eta_p^2 = .066$, (2) the bad director to have knowingly caused the damages of the suppliers to a larger extent than the good director (M = 5.70, SD = 1.13 vs. M = 4.80, SD = 1.31), $F(1,235) = 32.48$, $p < .00$, $\eta_p^2 = .121$, and (3) the bad director to have acted more recklessly in damaging the suppliers than the good

director (M = 6.30, SD = 0.88 vs. M = 5.95, SD = 1.05), $F(1,235) = 7.77$, $p = .006$, $\eta_p^2 = .032$). Figure 6.5 provides a visual presentation of the findings.

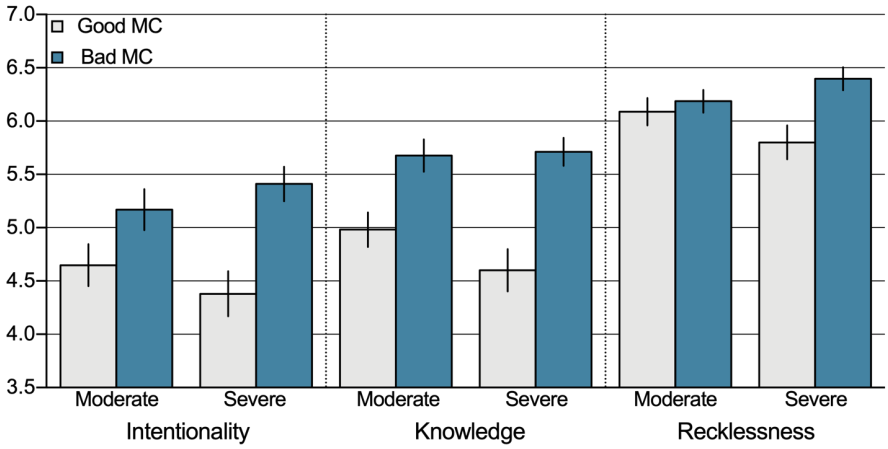

Figure 6.5. Mean scores for the three mental states, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

Also replicating Study 1, for the perceived likelihood of failure an effect was found for moral character $F(1,235) = 7.81$, $p = .006$, $\eta_p^2 = .032$, but not for outcome severity, $F(1,235) = 1.29$, $p = .257$, $\eta_p^2 = .005$, nor for the interaction between the two factors, $F(1,235) = 1.37$, $p = .243$, $\eta_p^2 = .006$. Participants again considered the likelihood of failure to be higher in the bad moral character condition than in the good moral character (M = 70.02, SD = 25.21 vs. M = 61.58, SD = 22.97).

Again replicating Study 1, an effect of moral character was found for punishment $F(1,235) = 14.43$, $p < .001$, $\eta_p^2 = .058$, but not for outcome severity, $F(1,235) = 1.45$, $p = .230$, $\eta_p^2 = .006$, nor for the interaction effect, $F(1,235) = 1.01$, $p = .317$, $\eta_p^2 = .004$. Participants assigned a higher percentage in the case with the morally bad director than with the morally good director (M = 78.88, SD = 24.36 vs. M = 66.51, SD = 24.47). Figure 6.6 visually presents the data for likelihood of failure and punishment.

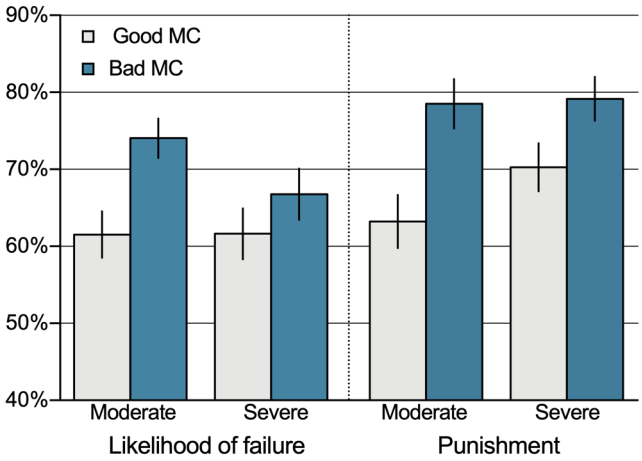

Figure 6.6. Mean scores for the perceived likelihood of failure and punishment, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

Finally, also replicating Study 1, a MANOVA with all three moral judgments as dependent variables returned a significant effect both for moral character, $F(3,233) = 6.84$, $p < .001$, $\eta_p^2 = .081$, as well as for outcome severity, $F(3,233) = 6.55$, $p < .001$, $\eta_p^2 = .078$, but not for the interaction, $F(3,233) = 2.40$, $p = .069$, $\eta_p^2 = .030$. Univariate analyses showed that the director's actions were judged as more wrong ($M = 6.05$, $SD = 1.24$ vs. $M = 5.36$, $SD = 1.37$), $F(1,235) = 15.49$, $p < .001$, $\eta_p^2 = .062$, more blameworthy ($M = 6.06$, $SD = 1.23$ vs. $M = 5.42$, $SD = 1.30$), $F(1,235) = 15.32$, $p < .001$, $\eta_p^2 = .047$, and more impermissible ($M = 5.52$, $SD = 1.51$ vs. $M = 4.85$, $SD = 1.66$), $F(1,235) = 11.65$, $p < .001$, $\eta_p^2 = .047$, when the case involved a morally bad director compared to a morally good director. In case of the severe outcome, the director's actions were judged as more wrong ($M = 5.93$, $SD = 1.28$ vs. $M = 5.55$, $SD = 1.37$), $F(1,235) = 4.00$, $p = .047$, $\eta_p^2 = .017$, but at the same time less impermissible (albeit not statistically significant strictly speaking) when the case had a severely bad outcome compared to a moderately bad outcome, ($M = 5.07$, $SD = 1.77$ vs. $M = 5.38$, $SD = 1.42$), $F(1,235) = 3.33$, $p = .069$, $\eta_p^2 = .014$. For the blameworthiness of the director, we found no effect of outcome severity, $F(1,235) = 0.59$, $p = .445$, $\eta_p^2 = .002$. See Figure 6.7 for a visual representation of these findings.

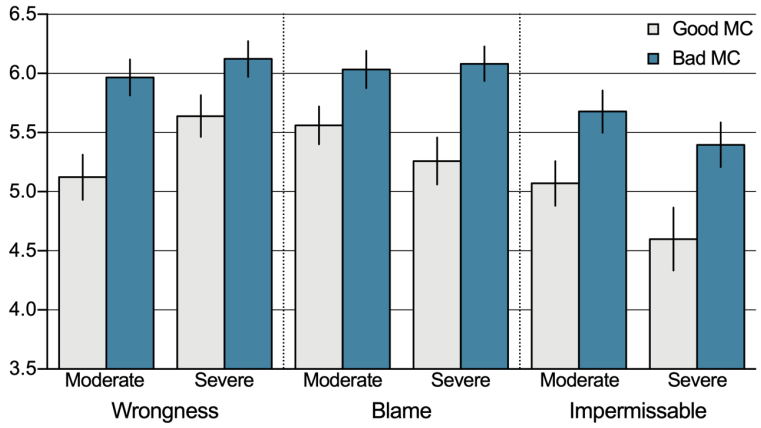

Figure 6.7. Mean scores for the three moral judgments, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

6.3.2 Discussion

Study 2 largely replicated Study 1, providing further support for the notion that moral character inferences have a strong effect on mental state ascriptions, perceived likelihood of failure, punishment, and moral judgments, and that moral character effects even seem to trump outcome effects. Despite the fact that in Study 2 it was explicitly stated the director himself believed the chance of failing and thus harming his suppliers was very large, participants were still affected by moral character inferences such that they were more likely to indicate the director intentionally damaged his suppliers, that he knew he would damage his suppliers, and that he knew there was a substantial risk he would damage the suppliers. Whereas Study 1 found effects of outcome severity for both wrongness and blame in the expected direction, Study 2 only found an effect for wrongness.

6.4 Study 3

A few limitations of Studies 1 and 2 remain that we addressed in Study 3. First, Study 1 and 2 manipulated the director's moral character using the same storyline (i.e., the director being pro or anti animal rights, see Appendix 6.1), while also referring to the director's reputation within international corporate circles. The latter aspect might affect people's judgments regarding the director's competence which might then affect the perceived likelihood of failure. Also, the fact that the same storyline was used makes it possible that the effects

are (to an extent) dependent on this particular storyline. Therefore, Study 3 addressed these issues by no longer referring to the director's reputation within his professional network and by using three different versions of the character manipulation (see Appendix 6.2; each participant received only one of three versions).

Second, the fact that the questions pertaining to the director's mental state were presented one by one on separate screens might somehow have affected the results. It has for example been argued that the Knobe effect can be explained by people's pragmatic use of language, as mental state ascriptions can be used to signal disapproval or blame, even though people might not genuinely believe an actor intentionally brought about a certain effect (e.g., Adams & Steadman, 2004b, 2004a). Therefore, in Study 3 we presented the mental state questions on the same screen as this might induce more reflective responses.

Finally, Studies 1 and 2 both used Dutch samples and it therefore remains unknown to what extent the observed effects are limited to this specific population. For Study 3, therefore, participants were recruited using Amazon's Mechanical Turk to investigate the effects in a predominantly US sample.

Similar to Study 1, the director's subjective likelihood of failure was stated as being very small and we therefore again included the negligence measure. In contrast to Studies 1 and 2, Study 3 only included blame as a moral judgment and omitted the measures for wrongness and permissibility, as blame was the key variable of interest (of the moral judgments). Also, in an attempt to elicit stronger outcome effects, the severe outcome in Study 3 was made more severe by stating: "Many [employees] even stated that losing their job meant they could not take proper care of their children anymore."

Only the data of participants who passed three simple comprehension questions (related to the case) was recorded. In total, 423 participants were recruited to ensure sufficient statistical power. Based on the same exclusion criteria used in Studies 1 and 2, 177 participants were excluded from analyses leaving a final sample of 246 participants. The mean age of the final sample was 39.4 (SD = 12.1) and 56.1% were male.

6.4.1 Results

The manipulation check showed that participants rated the morally bad director as morally worse than the morally good director (M = 6.61, SD = 0.70 vs. M = 3.88, SD = 1.49), $F(1,240) = 338.26$, $p < .001$, $\eta_p^2 = .585$. This was independent from the version of the character manipulation that was used, as indicated by a non-significant interaction $F(2,240) = 0.30$, $p = .740$, $\eta_p^2 = .003$.

A MANOVA with the four mental states as dependent variables and moral character and outcome severity as factors returned a significant effect for moral character, $F(4,239) = 16.38$, $p < .001$, $\eta_p^2 = .215$, and a near statistically signi-

ficant effect for outcome severity, $F(4,239) = 2.39$, $p = .051$, $\eta_p^2 = .039$. No significant interaction was found, $F(4,239) = .58$, $p = .680$, $\eta_p^2 = .010$. Subsequent univariate analyses showed that the morally bad director was judged as having acted more intentionally ($M = 4.48$, $SD = 1.54$ vs. $M = 3.13$, $SD = 1.69$), $F(1,242) = 43.00$, $p < .001$, $\eta_p^2 = .151$, knowingly ($M = 4.47$, $SD = 1.65$ vs. $M = 3.25$, $SD = 1.64$), $F(1,242) = 35.97$, $p < .001$, $\eta_p^2 = .129$, and recklessly ($M = 6.10$, $SD = 1.07$ vs. $M = 5.07$, $SD = 1.43$), $F(1,242) = 43.52$, $p < .001$, $\eta_p^2 = .152$, than the morally good director. No effect was found for negligence, $F(1,242) = 2.63$, $p = .106$, $\eta_p^2 = .011$. For outcome severity, the director was judged as having acted more knowingly ($M = 4.03$, $SD = 1.78$ vs. $M = 3.66$, $SD = 1.72$), $F(1,242) = 4.97$, $p = .027$, $\eta_p^2 = .020$ and recklessly ($M = 5.72$, $SD = 1.28$ vs. $M = 5.42$, $SD = 1.44$), $F(1,242) = 5.58$, $p = .019$, $\eta_p^2 = .023$, in case of the severe outcome relative to the moderate outcome (see Figure 6.8).

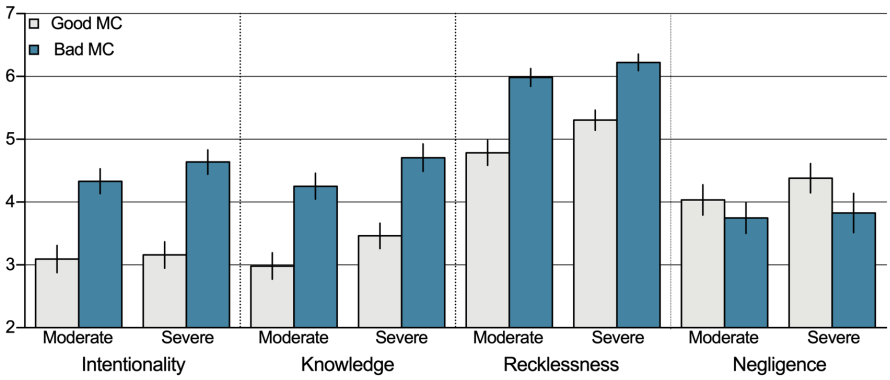

Figure 6.8. Mean scores for the four mental states, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

For the perceived likelihood of failure, an effect was found for both moral character, $F(1,242) = 43.66$, $p < .001$, $\eta_p^2 = .153$, and outcome severity, $F(1,242) = 8.16$, $p = .005$, $\eta_p^2 = .033$, but not for the interaction between the two, $F(1,242) = 0.81$, $p = .369$, $\eta_p^2 = .003$. The likelihood of failure was perceived to be higher in case of the morally bad director versus the morally good director ($M = 77.55$, $SD = 14.73$ vs. $M = 62.84$, $SD = 21.19$), as well as for the severe outcome relative to the moderate outcome ($M = 72.74$, $SD = 17.90$ vs. $M = 67.23$, $SD = 21.15$) (see Figure 6.9). For punishment, the same pattern was observed. A larger percentage of damages to be paid by the director was found for participants in the bad moral character condition relative to the good moral character condition ($M = 83.05$, $SD = 19.24$ vs. $M = 66.19$, $SD = 29.14$), $F(1,242) = 31.33$, $p < .001$, $\eta_p^2 = .115$, as well as for the severe outcome relative to the

moderate outcome (M = 77.19, SD = 22.69 vs. M = 71.60, SD = 29.16), $F(1,242)$ = 4.70, p = .031, η_p^2 = .019 (see Figure 6.9).

Finally, participants considered the morally bad director to be more blameworthy than the morally good director (M = 6.55, SD = 0.82 vs. M = 5.73, SD = 1.21), $F(1,242)$ = 40.04, p < .001, η_p^2 = .142. No effect was found for outcome severity $F(1,242)$ = 2.63, p = .106, η_p^2 = .011, nor was there an interaction effect, $F(1,242)$ = 0.28, p = .599, η_p^2 = .001.

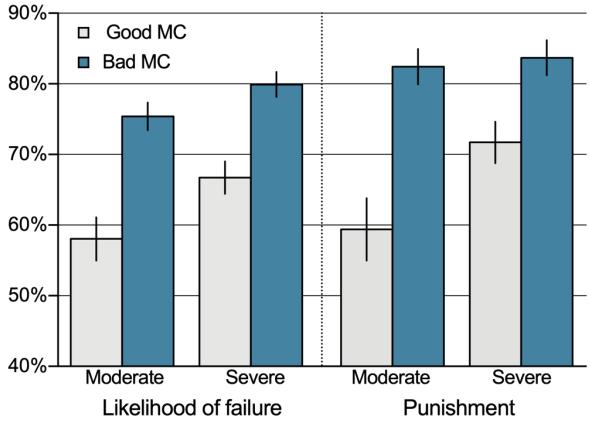

Figure 6.9. Mean scores for the perceived likelihood of failure and punishment, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

6.4.2 Discussion

Using a non-Dutch sample, while presenting the mental state questions on the same screen, and while using different moral character manipulations, the results from Study 3 largely replicated those from Studies 1 and 2. Specifically, moral character again had a significant effect on people's mental state attributions, their perceived likelihood of failure, as well as punishment and blame attributions. In Study 3, outcome severity seemed to have a stronger effect (albeit still much weaker than moral character) than observed in Studies 1 and 2. Specifically, in Study 3 we found that outcome severity affected participants' mental state attributions for knowledge and recklessness, as well as their perceived likelihood of failure and punishment attributions. In contrast, previously outcome severity only affected wrongness judgments (Study 1 and 2, not included in Study 3) and blame attributions (Study 1, no effect for blame in Study 3).

Hence, Study 3 provided further support for the notion that moral character inferences affect mental state attributions, as well as other legally relevant

judgments. Further, the findings from Study 3 suggest that outcome severity can in some cases have an effect independent from the effect stemming from moral character inferences.

6.5 Study 4

Ultimately, we wanted to compare a sample of legal professionals (in Study 5) with a sample of lay people. However, in Study 3 some information was included in the case that might have affected judgments of legal professionals in a way that would from a legal standpoint be justified, but that would not be related to moral character inferences per se. Specifically, in the version of the case with the morally bad director, the director states "I don't care at all about our suppliers, all I care about is making my own company profitable again". Legal professionals might infer from such a statement that the director did not take due care to protect his suppliers by for example conducting a proper investigation into the chances of being able to save his company or by calling in outside advisors for a second opinion. A lack of such measures might provide ground for a liability claim and therefore might have affected punishment attributions in the sample of legal professionals. Therefore, in Study 4 (and in Study 5) the case included the following paragraph (see Appendix 6.3 for the full case):

> When pressed by his advisor on the chances of survival, Devin said: 'Our CFO and external consultants have conducted careful analyses of our financial situation and also thoroughly analysed the turnaround plan. We are confident that there is a good chance we can save Yonos from bankruptcy. We believe that we have done everything we can to weigh all factors and consider it necessary to move forward with the turnaround plan. Disclosing our financial problems to our suppliers poses too great a risk. There is of course a possibility that our turnaround plan will fail and that we cannot pay our suppliers due to bankruptcy, but we consider the chance this will happen to be small.'

By including this paragraph, it is made clear the director took due care in relation to protecting his suppliers. Any effects of moral character are therefore unlikely to be due to assumptions about the director's precautionary measures. Participants might still feel the director should have informed his suppliers prior to commencing with his turnaround plan. Study 4 included this normative question as an additional dependent variable. Specifically, participants were asked to indicate on a 7-point Likert scale ranging from (1) "strongly disagree" to (7) "strongly agree" to what extent they agreed with the following statement: "The director should have informed his suppliers about the company's financial problems when placing the orders".

The methods and analysis plan for Study 4 were preregistered. A larger sample size was obtained relative to Study 1-3 to ensure sufficient power to

detect smaller effects. Data was collected for 438 participants using Amazon's MTurk. After excluding participants based on the same criteria as in Study 1-3, a final sample of 306 participants was obtained. The mean age was 40.5 (SD = 12.2), 51% were female, and 98.4% were native English speakers.

6.5.1 Results

The manipulation check showed that participants considered the morally bad director to have a morally worse character than the morally good director (M = 6.48, SD = 0.88 vs. M = 4.09, SD = 1.52), $F(1,300) = 274.43$, $p < .001$, $\eta_p^2 = .478$, and this effect was independent of the version of the manipulation that was used, as indicated by a non-significant interaction effect, $F(2,300) = 1.73$, $p = .180$, $\eta_p^2 = .011$

A MANOVA for the mental state variables returned a significant effect of moral character, $F(4,299) = 18.34$, $p < .001$, $\eta_p^2 = .197$, but not for outcome severity, $F(4,299) = 2.09$, $p = .082$, $\eta_p^2 = .027$, nor for the interaction, $F(4,299) = 1.02$, $p = .398$, $\eta_p^2 = .013$. Univariate analyses showed that the morally bad director was judged as having acted more intentionally (M = 4.45, SD = 1.69 vs. M = 2.86, SD = 1.60), $F(1,302) = 65.90$, $p < .001$, $\eta_p^2 = .179$, knowingly (M = 4.13, SD = 1.68 vs. M = 2.92, SD = 1.44), $F(1,302) = 41.90$, $p < .001$, $\eta_p^2 = .122$, and recklessly (M = 5.91, SD = 1.29 vs. M = 5.21, SD = 1.54), $F(1,302) = 16.26$, $p < .001$, $\eta_p^2 = .051$, but not more negligently, $F(1,302) = 0.18$, $p = .721$, $\eta_p^2 = .001$, than the morally good director. Despite the lack of a significant multivariate effect of outcome severity, a significant univariate effect of outcome severity was found for intentionality, such that the director was judged as having acted more intentionally in case of the severe outcome relative to the moderate outcome, (M = 4.00, SD = 1.79 vs. M = 3.28, SD = 1.80), $F(1,302) = 6.90$, $p = .009$, $\eta_p^2 = .022$ (see Figure 6.10).

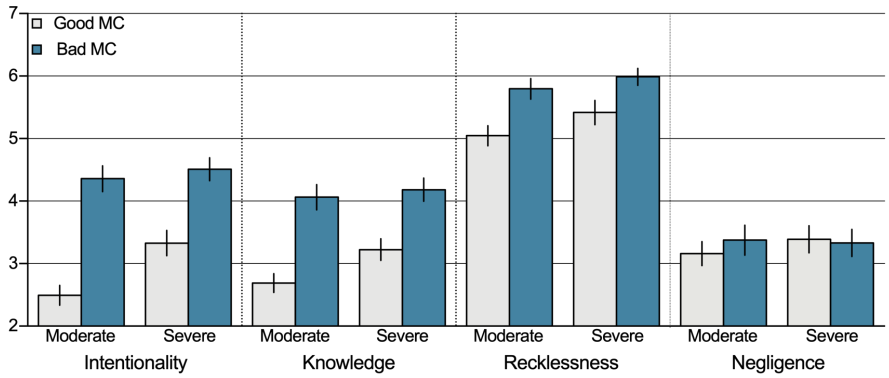

Figure 6.10. Mean scores for the four mental states, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

The perceived likelihood of failure was scored higher in the bad moral character condition relative to the good moral character condition, (M = 70.35, SD = 20.66 vs. M = 60.94, SD = 21.42), $F(1,302) = 13.16$, $p < .001$, $\eta_p^2 = .042$. The perceived likelihood of failure was also higher in case of the severe outcome than in case of the moderate outcome (M = 68.52, SD = 21.07 vs. M = 62.62, SD = 21.66), albeit not statistically significant, $F(1,302) = 3.63$, $p = .058$, $\eta_p^2 = .012$. No interaction effect between moral character and outcome severity was found $F(1,302) = 0.96$, $p = .329$, $\eta_p^2 = .003$. Higher punishment ratings were found for the morally bad director compared to the morally good director (M = 78.47, SD = 26.09 vs. M = 66.44, SD = 28.57), $F(1,302) = 15.40$, $p < .001$, $\eta_p^2 = .049$. No significant effect was found for outcome severity, $F(1,302) = 0.49$, $p = .486$, $\eta_p^2 = .002$, nor for the interaction, $F(1,302) = 1.88$, $p = .172$, $\eta_p^2 = .006$ (see Figure 6.11).

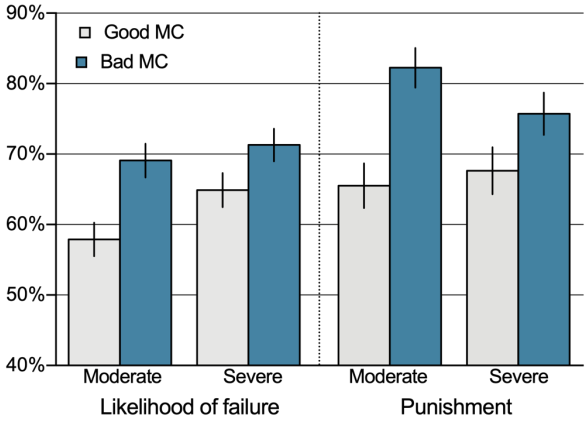

Figure 6.11. Mean scores for the perceived likelihood of failure and punishment, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

Higher blame attributions were found for the morally bad director versus the morally good director (M = 6.51, SD = 0.94 vs. M = 5.72, SD = 1.31), $F(1,302) = 35.64$, $p < .001$, $\eta_p^2 = .106$. No effect of outcome severity, $F(1,302) = 0.04$, $p = .846$, $\eta_p^2 = .000$, nor an interaction effect was found, $F(1,302) = 0.29$, $p = .593$, $\eta_p^2 = .001$. Regarding the normative question (i.e., "the director should have informed his suppliers about the company's financial problems when placing the orders"), higher scores were found for the morally bad director than for the morally good director (M = 6.25, SD = 1.04 vs. M = 5.70, SD = 1.34), $F(1,302) = 15.38$, $p < .001$, $\eta_p^2 = .048$. No effect of outcome severity, $F(1,302) = 0.22$, $p = .643$, $\eta_p^2 = .001$, nor an interaction effect was found, $F(1,302) = 3.34$, $p = .068$, $\eta_p^2 = .011$ (see Figure 6.12).

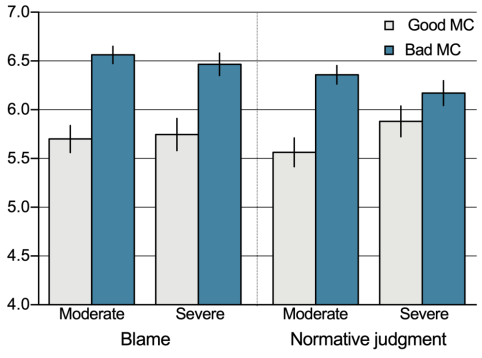

Figure 6.12. Mean scores for blame judgments and the normative judgments regarding what the director *should* have done, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

6.5.2 Discussion

Study 4 fully replicated Study 3 for the effects of moral character, but some differences were observed for outcome severity. Whereas in Study 3 we found outcome effects for the mental states knowledge and recklessness, Study 4 only found an effect for intentionality. Additionally, whereas Study 3 found an effect of outcome severity for the perceived likelihood of failure and punishment attributions, no such effects were found in Study 4. Hence, across the studies thus far the results for outcome severity have been somewhat inconsistent and generally small. So far it seems therefore that moral character inferences have a significantly more pronounced and consistent effect than outcome severity.

6.6 Study 5

Study 5 was identical to Study 4 and we aimed to compare the lay people sample used in Study 4 with a sample of legal professionals. The methods and analyses plan for Study 5 were preregistered. Participants were professionals specialized in the areas of insolvency law, business restructuring and/or recovery and were all members of INSOL International, which is a world-wide federation of national associations of professionals who specialize in turnaround and insolvency. Participants were approached via e-mail with an invitation to participate in our study. The participants' e-mail addresses were obtained from INSOL International's online membership directory. We initially aimed to recruit 350 participants, but in an attempt to reach 240 legal professionals from the UK specifically, we recruited 425 participants in total. In the end, we managed to recruit 124 legal professionals from the UK. After excluding 24 participants based on the preregistered exclusion criteria (same as in Studies 1-4), we obtained a final sample of 401 participants.

The mean age of this final sample was 48.4 (SD = 11.2) and 81.5% were male. Participants had on average 22.3 years of professional experience (SD = 10.8), and 85.8% indicated that investigating or deciding over directors' liability issues is part of their job.[2] In the final sample, 54 different nationalities were represented, with the top three consisting of the United Kingdom (29.4%), Australia (14.0%), and Canada (7.7%). Appendix 6.4 contains a complete overview of the nationalities represented in the survey.

2 Analyses with only participants who indicated to investigate or decide over directors' liability issues resulted in similar results and significance levels and therefore did not affect the conclusions.

6.6.1 Results

Despite not hitting the target sample size for UK participants, we checked whether this factor (UK vs. non-UK) interacted with moral character or outcome severity for any of the DVs. No significant interactions were found (all *p*s > .101) and we therefore do not distinguish between UK and non-UK legal professionals in further analyses.

The manipulation check showed that participants rated the morally bad director as having a worse character than the morally good director ($M = 5.79$, $SD = 1.16$ vs. $M = 3.30$, $SD = 1.36$), $F(1,395) = 384.29$, $p < .001$, $\eta_p^2 = .493$, and this was independent from the version used as indicated by a non-significant interaction, $F(2,395) = 0.20$, $p = .820$, $\eta_p^2 = .001$.

A MANOVA with the four mental states as DVs and moral character and outcome severity as factors returned a significant effect for moral character, $F(4,394) = 16.03$, $p < .001$, $\eta_p^2 = .140$, but not for outcome severity, $F(4,394) = 0.60$, $p = .663$, $\eta_p^2 = .006$, nor for the interaction, $F(4,394) = 0.65$, $p = .625$, $\eta_p^2 = .007$. Univariate analyses showed effects of moral character for intentionality ($M = 3.46$, $SD = 1.83$ vs. $M = 2.24$, $SD = 1.42$), $F(1,397) = 56.43$, $p < .001$, $\eta_p^2 = .124$, knowledge ($M = 3.11$, $SD = 1.61$ vs. $M = 2.28$, $SD = 1.22$), $F(1,397) = 32.21$, $p < .001$, $\eta_p^2 = .077$, and recklessness ($M = 5.33$, $SD = 1.65$ vs. $M = 4.42$, $SD = 1.85$), $F(1,397) = 26.72$, $p < .001$, $\eta_p^2 = .063$, but not for negligence, $F(1,397) = 0.74$, $p = .389$, $\eta_p^2 = .002$ (see Figure 6.13).

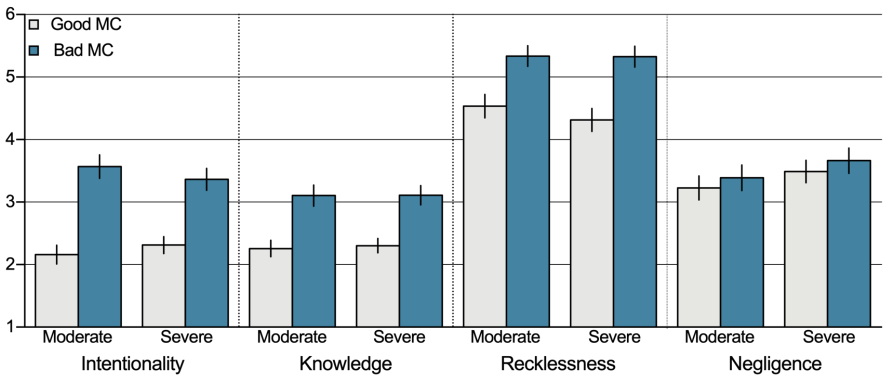

Figure 6.13. Mean scores for the four mental states, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character)

For the perceived likelihood of failure, an effect was found for moral character ($M = 56.52$, $SD = 19.60$ vs. $M = 47.17$, $SD = 20.63$), $F(1,397) = 22.19$, $p < .001$, $\eta_p^2 = .053$, but not for outcome severity, $F(1,397) = 0.01$, $p = .941$, $\eta_p^2 = .000$, nor for the interaction, $F(1,397) = 1.99$, $p = .159$, $\eta_p^2 = .005$. For punishment,

an effect was found for moral character (M = 45.73, SD = 37.44 vs. M = 34.55, SD = 38.63), $F(1,397) = 9.21$, $p = .003$, $\eta_p^2 = .023$, but not for outcome severity, $F(1,397) = .80$, $p = .371$, $\eta_p^2 = .002$, nor for the interaction, $F(1,397) = 3.04$, $p = .082$, $\eta_p^2 = .008$ (see Figure 6.14).

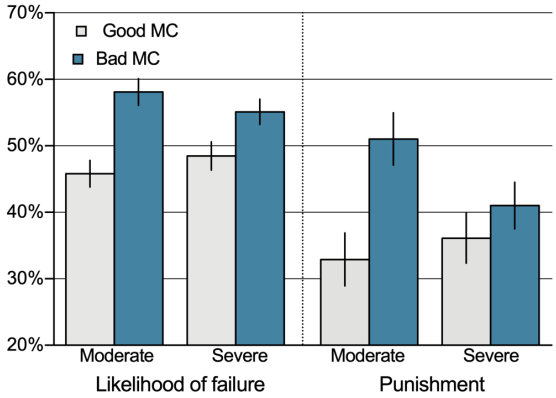

Figure 6.14. Mean scores for the perceived likelihood of failure and punishment, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

For blame, an effect was found for moral character (M = 5.27, SD = 1.62 vs. M = 4.41, SD = 1.83), $F(1,397) = 24.62$, $p < .001$, $\eta_p^2 = .058$, but not for outcome severity, $F(1,397) = 0.07$, $p = .798$, $\eta_p^2 = .000$, nor for the interaction, $F(1,397) = 0.62$, $p = .433$, $\eta_p^2 = .002$. For the normative question, the effect of moral character did not reach statistical significance (M = 4.36, SD = 1.71 vs. M = 4.05, SD = 1.81), $F(1,397) = 3.49$, $p = .063$, $\eta_p^2 = .009$, nor was there an effect of outcome severity, $F(1,397) = 0.33$, $p = .565$, $\eta_p^2 = .001$, or of the interaction, $F(1,397) = 2.02$, $p = .156$, $\eta_p^2 = .005$ (see Figure 6.15).

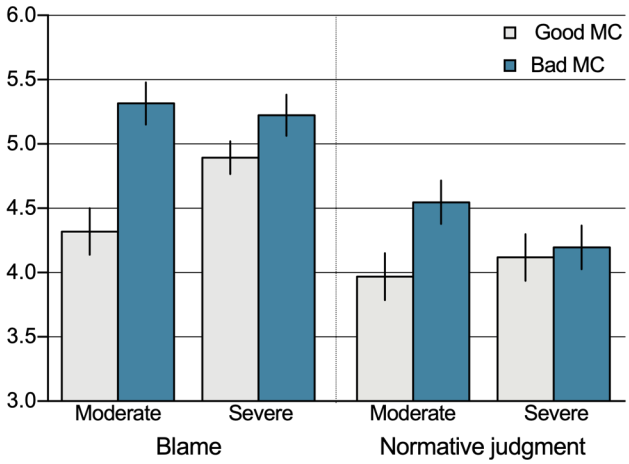

Figure 6.15. Mean scores for blame judgments and the normative judgments regarding what the director *should* have done, separated by outcome severity (moderate vs. severe) and moral character (Good MC = good moral character; Bad MC = bad moral character).

Exploratory analyses were conducted to directly compare lay people with legal professionals. We tested whether the population factor of lay people vs. legal professionals interacted with moral character or outcome severity for any of the DVs. Only a significant interaction was found between the population factor and outcome severity for intentionality, $F(1,699) = 4.26$, $p = .039$, $\eta_p^2 = .006$. Whereas lay people judged the director to have acted more intentionally in case of the severe outcome relative to the moderate outcome ($M = 4.00$, $SD = 1.79$ vs. $M = 3.28$, $SD = 1.80$), $F(1,302) = 6.90$, $p = .009$, $\eta_p^2 = .022$, no such effect of outcome severity in intentionality judgments was found for legal professionals ($M = 2.85$, SD = 1.70 vs. M = 2.85, SD = 1.81), $F(1,397) = 0.20$, $p = .891$, $\eta_p^2 = .000$. However, considering we did not find a significant multivariate interaction effect (i.e., with all four mental states as DVs) between the population factor and outcome severity, $F(4,700) = 1.98$, $p = .095$, $\eta_p^2 = .011$, the difference between the two populations in intentionality judgments due to outcome severity should be interpreted with caution. We did consistently find a main effect for the population factor across all DVs, apart from negligence, with higher scores for lay people than for legal professionals (see Table 6.1).

Table 6.1. Mean scores for lay people and legal professionals across the DVs, as well as significance tests for the mean differences.

	Lay people (N = 306)	Legal professionals (N = 401)			
DV	M (SD)	M (SD)	F(1,705)	p	η_p^2
Intentionality	3.65 (1.83)	2.85 (1.75)	34.45	< .001	.047
Knowledge	3.53 (1.68)	2.70 (1.49)	47.69	< .001	.063
Recklessness	5.56 (1.46)	4.88 (1.81)	28.47	< .001	.039
Negligence	3.30 (1.88)	3.45 (1.95)	0.99	.320	.001
Likelihood of failure	65.61 (21.53)	51.88 (20.63)	73.99	< .001	.095
Punishment	72.42 (27.98)	40.18 (38.40)	153.43	< .001	.179
Blame	6.11 (1.21)	4.84 (1.78)	114.86	< .001	.140
Normative judgment	5.97 (1.23)	4.20 (1.76)	224.41	< .001	.241

6.6.2 Discussion

Overall, Study 5 largely replicated Study 4 and the data therefore suggests that legal professionals are similarly affected by moral character inferences as lay people in their mental state attributions, as well as in their perceptions of the likelihood of failure, their punishment attributions, and their blame attributions. However, two differences between the two samples surfaced. First, intentionality judgments were affected by outcome severity in the sample of lay people, but not in the sample of legal professionals. Second, an effect of moral character was found for the normative judgment in lay people, but not in legal professionals. Interestingly, across all DVs (apart from negligence) we found lower scores for legal professionals than for lay people, suggesting that legal professionals were generally less punitive in their judgments.

6.7 General Discussion

The current paper put forward a novel account of the folk psychology of intentional action and set out to provide empirical support for its central role of moral character inferences in morally laden contexts. Also, the paper aimed to shed further light on the relationship between outcome effects and moral character inferences in mental state attributions as well as other legally relevant judgments (i.e., likelihood of failure, punishment, blame). Finally, an important contribution of the current paper is that we compared lay people with legal professionals to test whether the latter group can also be influenced by ir-

relevant character information when being presented with a detail rich case on a topic in which they are specialized.

Overall, we found overwhelming support for the notion that irrelevant information pertaining to an agent's moral character influences mental state attributions, perceptions of a company's outlook, as well as blame and punishment attributions. Such effects were found irrespective of the way in which the mental state questions were presented (i.e., separate versus jointly), the way in which moral character was manipulated (i.e., the 'story' that was used), small changes that were made to the case, and of the specific population under investigation (i.e., Dutch lay people, US lay people, legal professionals).

Effects of outcome severity were rather small and inconsistent. For blame attributions only an effect of outcome severity was found in Study 1. Wrongness judgments were affected by outcome severity in both Studies 1 and 2. In Study 3 we found outcome effects for the mental states knowledge and recklessness, as well as for the perceived likelihood of failure and punishment. In Study 4 we only found an effect of outcome for intentionality judgments and in Study 5 we did not find any effect of outcome severity. Moreover, we did not find any evidence for an interaction effect between moral character inferences and outcome severity. It seems therefore that when (irrelevant) character information is provided, outcome information has very little additive effect.

6.7.1 Theoretical Implications

The above findings help shed light on some discussions in the literature. First, concerning the debate on whether moral considerations affect mental state ascriptions, the studies presented here suggest that at least in morally laden contexts moral considerations do indeed affect mental state ascriptions, which goes against accounts that argue evaluative factors can be explained away by non-moral factors (e.g., Guglielmo & Malle, 2010a, 2010b; Machery, 2008; Wright & Bengson, 2009). In relation to the debate on whether evaluative considerations affecting mental state ascriptions represent a bias or whether such considerations are a constitutive component of intentional action (see for example Cova, 2016; Kneer & Bourgeois-Gironde, 2017), the present studies speak in favour of the former.

One category of bias accounts that has received some criticism is that focusing on the blameworthiness of an agent (e.g., Alicke, 2008; Nadelhoffer, 2004a, 2004b, 2006). A lot of weight has been given to a study among seven individuals with brain damage (i.e., impairment of emotional processing) for whom the typical asymmetry in intentionality judgments in the chairman scenario was observed nonetheless (Young & Cushman, 2006). This finding has been used to argue that blame processes cannot account for the Knobe effect. However, the current studies provide strong evidence that at least in

morally laden contexts the blameworthiness of an agent as derived from his/her moral character does affect mental state ascriptions as well as related judgments. Hence, even though it is possible that the necessity of affective processing might be limited, we probably should not throw out the proverbial baby with the bath water. The current research extends previous blame accounts by showing that the blameworthiness of the agent need not be derived from the action or (side-) effect under consideration, but is rather, ultimately, derived from inferences regarding the agent's moral character, which might be based on unrelated information sources.

Finally, Cova (2016) has listed four criteria that any adequate account of intentional action should meet, which our Moral Character Account does. First, Cova states: "A proper account should explain both the Knobe Effect and the Skill Effect." The Skill Effect refers to findings discussed in the introduction of this paper, showing that skill or control is a relevant factor for intentionality judgments in non-moral contexts (marksman shooting a target), but not (or at least much less so) in morally laden contexts (marksman killing his aunt). Our MCA's distinction between morally neutral and morally laden contexts that dictates which concept of intentionality is made salient (i.e., causal or moral responsibility, respectively), can perfectly account for both the Knobe Effect as well as the Skill Effect. Cova's second condition is: "A proper account should explain why asymmetries similar to the original Knobe Effect can be observed in cases involving no moral violation." We have argued that the sales case as put forward by Wright and Bengson (2009), on which this condition is based, does not actually constitute a morally neutral scenario and can be explained by moral coherence processes following from moral character judgments. Cova's third condition is: "A proper account should accommodate the fact that the agent's attitude towards a side-effect (whether he brings it about reluctantly, indifferently, or joyfully) has an impact on our ascriptions of intentionality." Our MCA incorporates agents' attitudes towards side effects in such a way that these speak to the agents' moral characters. Bringing about a bad side effect wholeheartedly signals a worse moral character than bringing about the same side effect reluctantly or regretfully. Cova's last condition is: "A proper account should account for the fact that norms seem able to drive asymmetries similar to the Knobe Effect independently of side-effects' valence." This condition was inspired by the racial identification law in Nazi Germany scenario and we have argued that the findings can be explained by our MCA's required proportionality between an agent's moral character and the blame that is implied by acting in a norm-violating manner. Even though the agent in that specific scenario fulfilled the requirements of a morally reprehensible law, the blame that would be implied by saying he intentionally did so would be disproportionate to the amount of blame the agent deserved based on his character.

In short, based on the criteria set out by Cova (2016), our MCA seems to be the most adequate account available. Based on meeting Cova's criteria as

well as the empirical support for our account provided in this paper, we consider our MCA to be of value for helping us understand the folk psychology of intentional action, while also being able to explain a host of related mental state ascriptions as well as other legally relevant judgments such as causality, freedom, foreseeability, etc.

6.7.2 Practical Implications

In addition to the above theoretical considerations, the findings should also concern legal practice. That is, given the evidence presented here of legal professionals being affected in their judgments by irrelevant information concerning an agent's moral character, the question should be to what extent this is problematic and what can be done to prevent such biased judgments. Even though lawyers are not the ultimate decision makers in legal cases, in many jurisdictions they nonetheless play a pivotal role and are sometimes the only ones who can bring a liability claim against a director in an insolvency proceeding. Hence, whereas previous work already voiced concerns regarding the impartiality of juries (consisting of lay-people)(Nadelhoffer, 2006), the current research suggest legal professionals might not be exempt.

Whereas previous research already found that judges are equally susceptible to the Knobe effect as lay people and that this most likely constitutes a bias (Kneer & Bourgeois-Gironde, 2017), the present research extends these findings by demonstrating that moral considerations can also affect legally relevant judgments other than intentionality, such as ascriptions of knowledge and recklessness, post-hoc perceptions regarding the likelihood of an adverse event occurring, as well as blame and punishment attributions. Regarding the perceived likelihood of failure as measured in the sample of legal professionals of the current research, we even found that moral character information can push likelihood perceptions from below the midpoint when the director had a good moral character (i.e., 47.2%) to above the midpoint when the director had a bad moral character (i.e. 56.5%). This is particularly relevant in alleged cases of wrongful trading as directors face liability when they continue to trade while it is 'more likely than not' that their company will become insolvent. Hence, perceptions of good moral character might tip the scale in a director's favour while a director's bad moral character might bias likelihood perceptions in such a way that in hindsight it will appear as if the company was facing insolvency and the director can thus be held liable for continuing to trade.

It is important to point out that character information might not only bias judgments in criminal proceedings where character evidence is admissible and actually a common element of court hearings in the US (see Federal Rules of Evidence; Saltzburg, Martin, & Capra, 1998), but also in civil cases such as is used in the present studies. The current findings therefore highlight and further extend the concerns that have been raised regarding the admissibility

of character information in criminal proceedings (e.g., Hunt & Budesheim, 2004; Maeder & Hunt, 2011; Uviller, 1982) to civil cases. Despite that character evidence not playing a similar role in civil cases as it currently does in criminal proceedings, in civil cases too there are many ways in which lawyers and judges are exposed to information that directly speaks to a tortfeasor's character.

Parallel to the discussion whether character information should be admissible or not in criminal proceedings, we encourage legal scholars to engage with psychologists and philosophers to think of ways to limit the biasing effect character information might have.

6.7.3 Limitations and Future Research

A few limitations and open questions remain. First, to get a better understanding of how moral character information and outcome information might independently or jointly affect mental state attributions and related judgments, more research is needed. In the current studies we only included two bad outcomes which differed in severity. We consider it worthwhile to test the relationship between character and outcome when using outcomes of opposite valance (i.e., good and bad outcomes).

Second, the absence of any strong or consistent outcome effect might have been due to the participants not really considering the outcomes to be that bad at all. Possibly, people care more about the environment being harmed in some way than about businesses going bankrupt. It would therefore be worthwhile to further investigate outcome effects in relation to moral character effects while using outcomes that trigger a stronger emotional response.

Third, in the present studies it turned out to be difficult to convince the participants of the director's good moral character, as indicated by scores around the midpoint of the scale in the good moral character condition. Participants might still have used the director's actions (jeopardizing the company's creditors) to infer that his character was less than ideal. Alternatively, participants might have had a strong bias against corporate directors such that even engaging in very charitable behaviour, being a loyal family man and good father, or being an outspoken environmentalist might not have been sufficient to wash away people's preconceptions about corporate directors. The fact that participants did not consider the morally good director to be a very good person means our findings are perhaps on the conservative end and it would therefore be worthwhile to study the hypotheses in different settings using different study materials to better manipulate agents' moral character.

Finally, even though our Moral Character Account of the folk psychology of intentional action suggests that moral character effects only come into play in morally laden contexts, it would be worthwhile to test under which circum-

stances such character effects might also play a role in morally neutral contexts. It might be that there is some cross-over from moral responsibility to causal responsibility when the (side-)effect is something one can get credit for. For example, in the shooting competition scenario, people might be reluctant to say the highly skilled marksman intentionally hit the target when the marksman is described as having a very bad moral character, as achieving something difficult that requires a lot of skill is typically something worthy of praise and people are probably reluctant to praise a morally bad marksman.

6.7.4 Conclusion

In this paper we have put forward a novel account of the folk psychology of intentional action and have argued that it can explain all existing data demonstrating asymmetries in mental state ascriptions, while also being able to account for data related to judgments such as causality, freedom, and foreseeability. The central tenet of our account is the role of moral character inferences in morally laden context. We have provided evidence for the notion that, even when keeping many factors that have previously been argued to explain intentionality judgments fixed, moral character inferences based on irrelevant information can still bias mental state ascriptions and other legally relevant judgments. Apart from advancing theorizing around the folk psychology of intentional action, the findings clearly pose a problem for legal practice as biased judgments stemming from character information threaten the notion of a fair trial. Directions for future research have been suggested to better understand the precise role that character inferences might have on important judgments.