The Role of Moral Character in Blaming Failure to Help

HIROZAWA Paula Yumi

Table of Contents

Abstract

In the present Ph.D. thesis, I explored the role of moral character in assigning blame for failure to help. In the general principal of law, blame is assigned to a crime based on its objective element, i.e., the action or inaction itself, and the subjective element, i.e., the mental state of the criminal. Moral psychological research sustains the relevance of these two criteria in the psychological processing of blame. However, recent research has also explored another important element in assigning blame: the moral character of the agent. Blame has been predominantly studied from the perspective of blameworthy wrongdoings. The present thesis aims to explore the roles of action-inaction and mental state information for assignment of blame in situations of failure to help, instead of wrongdoings, with the focus on the moral character explanation. In three sets of studies, I argue that the moral character of the agent is a determinant process through which individuals make judgments of blame for failure to help. In Study 1, I show evidence that choosing to omit help (vs. taking action to help) affects blame for failure to help in part due to underlying perceptions of moral character. In Study 2, I show that mental states such as intentions are informative of character, especially when the mental state is negative. In Study 3, I provide evidence that mental state information affects blame for failure to help due to inferences of moral character. Overall, the present thesis showed that moral character is an important explanation to the effect of both action-inaction and mental state information on blame for failure to help.

Acknowledgment

A big-hearted appreciation to Prof. Minoru Karasawa, who walked me through this academic experience with lots of patience and inspirational support. I am also very grateful for the support of my labmates, the faculty members, the Nagoya University staff, and the financial support of the Ministry of Education, Culture, Sport, Science, and Technology (MEXT) of the Japanese Government.

Chapter 1 – Theories and concepts

1.1 Blaming failure to help

The golden rule of morality is that one should treat others as they would like to be treated themselves. This rule implies that individuals should be kind to one another and that there should be a level of reciprocity in individual's minds and behaviors. The origins of helping behavior likely stem from the human evolution through cooperation with kin and reciprocation with people unrelated to them (Barret et al., 2002; Boster et al., 2001). Indeed, most psychological theories of morality argue that cooperation is the basis of all morality. Helping others is a fundamental moral value reinforced across a wide range of societies (Graham et al., 2013; Shweder et al., 1997). From a nativist viewpoint, the current complex psychological system of caring for others is based on the sensory signs of suffering of the offspring, which evolved to a complex motivation to invest in others' needs and avoid pain and suffering for oneself and others (Graham et al., 2013; Haidt & Joseph, 2008). An approach to morality as cooperation (Curry et al., 2019) argues that morality is the result of biological and cultural mechanisms developed across the evolution of humanity to provide the motivation for cooperative behavior. The bottom line is that the abilities to cooperate, help and care for others are essential for modern societies and fundamental to moral judgments.

Naturally, it follows that failing to cooperate, help or care for others should stem blame. Curiously, blame for failing to help has not been nearly as studied as blame for committing wrongdoing. Most of the research that has investigated blame for failure to help has focused on the differences between moral judgments of omissions and commissions (Baron & Ritov, 2004; Ritov & Baron, 1999; Spranca et al, 1991). The so-called omission effect refers to the finding that people assign harsher judgments to committed negative acts (e.g., killing) compared to omissions (e.g., not stopping one from killing). This sort of manipulation is certainly appropriate to allow an

investigation of the psychological processes underlying harsher judgments for increasing "acted out behavior," however it is not informative of the processes that lead people to blame omissions at all. That is, although there is large evidence that "harming*"* is more blameworthy than "not helping," it's still unknown how "not helping" is also blameworthy by itself. Moreover, there are situations in which omissions can be as blameworthy as commissions. In a study by DeScioli et al. (2011), an agent who omitted help by choosing to push a button that had no consequence in helping someone instead of one that would (i.e., "transparent" omission) was punished more than an agent who did not push any button (i.e., "opaque" omission). Moreover, the transparent omission was judged as harshly as pushing a button that directly caused the victim's death (i.e., commission). This effect was replicated even when controlling for perceived intentionality and causality. When portrayed as a conscious choice, omitting help can be as relevant to elicit blame as intuitively obvious sources of blame, such as commission.

A greater understanding of blame for commission and omission can be derived from Janoff-Bulman et al. (2009)'s exploration of what they call proscriptive and prescriptive morality. They approach morality from a self-regulation perspective, in which individuals are motivated to either avoid negative desires and outcomes (i.e., proscriptive morality), or to approach positive desires and outcomes (i.e., prescriptive morality). *Proscriptive morality* entails the inhibition of harmful behavior, or what *one shouldn't do,* such as the classic moral codes of no harming and no unfair treatment to others. On the other hand, *prescriptive morality* concerns promoting well-being and helping others, that is, what one *should* do, such as being generous, helpful, industrious, and hard-working. Violations of the proscriptive morality mean that the individual failed to inhibit their harmful behavior, whereas violations of the prescriptive morality suggest that the individual failed to activate their helpful behavior. This differentiation is important because, for the first time, Janoff-Bulman et al. (2009) clearly distinguish between two categories of immoral acts, that

of *doing harm* and that of *not helping*; *killing* and *not saving;* or *stealing* and *not donating*, for examples (see Table 1).

Importantly, Janoff-Bulman et al. (2009) also investigated for a potential positive-negative asymmetry between these two types of morality. It is well-known that several psychological processes (e.g., attention, memory, and person perception) show a negativity bias, that is, negative events tend to have a greater weight compared to positive events (Kensinger et al., 2006; Klein, 1991; Öhman et al., 2001; Peeters & Czapinsky, 1990; Pizarro et al., 2003; Reeder & Brewer, 1979; Rozin & Royzman, 2001). Their results were consistent with the negativity bias hypothesis for the proscriptive realm. For instance, they compared how much participants disapproved a target's either immoral proscriptive behaviors (e.g., going into debt by buying a TV) or immoral prescriptive behaviors (e.g., pass by a homeless man). They found that there was greater disapproval for committing negative acts than omitting positive ones. This is a very important finding to provide a framework to the omission effect – people seem to give more moral emphasis on behavior that should be avoided (i.e., proscriptive behavior) compared to the commendatory (or prescriptive) behaviors (Janoff-Bulman et al., 2009, Study 5).

Janoff-Bulman et al. (2009) also showed an interesting finding in their Study 3. They compared proscriptive and prescriptive adjectives (e.g, be aggressive; be kind, respectively) and verbs (e.g., steal; donate to charity, respectively) and measured the extent to which participants believed a person should or should not endorse in those actions/have those qualities and the extent to which they believed such actions/qualities are a personal choice. The results indicated that both proscriptive and prescriptive moral actions/qualities were equally recommended, yet prescriptive morality was perceived as being more of a personal choice than proscriptive morality. Moreover, when considering the adjectives versus verbs comparison, participants believed that an individual should *be* more moral prescriptively (e.g., be kind) than *not be*

immoral proscriptively (e.g., be aggressive), yet they should not engage in morally proscriptive behaviors (e.g., steal) to a greater extent than they should engage in morally prescriptive behaviors (e.g., help). One important conclusion, in consonance with their findings from Study 5, is that *not harming* seems to be a more condemnatory behavior than *helping*. Another important finding is that people also commend others to have a *moral character* (e.g., being kind, generous, and fair) to a greater extent than they commend one to simply *not be immoral*. That is, socially speaking, people believe that "others should be *moral* individuals and *not harm* others" more than they believe that "people should *not be immoral individuals* and *should help others*". This hint of the relevance of the moral character for prescriptive morality judgments will be a central point of this thesis.

In a reflection of their work, Janoff-Bulman et al. (2009) point out that, in a societal level, proscriptive morality is regulated by legal systems, which dictate what behaviors are prohibited and which punish the transgressors. On the other hand, there are fewer institutional tools to punish failure to help. For instance, in common law, people are not obligated to help others, unless they have a legal duty to do so, at least in Western cultures. Indeed, few states in the United States adopt Good Samaritan laws (Dressler, 2012). Hence, prescriptive morality is mostly regulated by social norms, based on expectations and social obligations of individuals in their communities. Evidence for this is shown by Buckwalter & Turri (2015). In their study, participants recognized that a healthy man who sees a child drowning in a pond does not have a legal obligation to help the child, yet they still perceived the man as *morally* obligated to help.

Table 1. Two Systems of Moral Regulation (based on Janoff-Bulmann et al. (2009)

	Proscriptive morality	Prescriptive morality
Motivation	Avoidance of negative desires and outcomes	Approach of positive desires and outcomes
Behavioral pattern	Inhibition of harmful behavior ("what one should not morally do")	Promotion of well-being and helping behavior ("what one should morally do")
Characteristics	Focused on transgressions; mandatory; condemnatory	Focused on "good deeds"; discretionary ("personal choice"); commendatory
Moral violation	Failure to inhibit harmful behavior (e.g., commission, such as failure to inhibit aggressive behavior); punishment for this violation is regulated by formal, legal institutions (e.g., imprisonment)	Failure to engage in prosocial behavior (e.g., omission, such as failure to help one in need); punishment for this violation occurs in the realm of social norms, based on social expectations (e.g., gossiping, defamation, ostracism)

Punishment and regulation of behaviors in the realm of prescriptive immorality likely occur by social mechanisms such as gossiping, reputation defamation, and social ostracism. Consider, for instance, the unfortunate case of Jamel Dunn, a man who was filmed by laughing teenagers as he drowned in a pond in Florida, United States, 2017. Legally speaking, the teenagers could not be held liable for failing to help the man, as they do not have a legal duty to rescue. However, it is possible to infer a high motivation to punish these teenagers in the following excerpts of an interview to CNN (Valencia & Sayers, 2017):

> "We don't have anything criminal resulting from that incident," Martinez said. "Our detectives were trying to get potentially if a negligence law could apply. The state attorney advises it doesn't meet standard for a criminal charge."

> "We are deeply saddened and shocked at both the manner in which Mr. Dunn lost his life and the actions of the witnesses to this tragedy," the state attorney's office said in a statement. "While the incident depicted on the recording does not give rise to sufficient evidence to support a criminal prosecution under Florida statutes, we can find no moral justification for either the behavior of persons heard on the recording or the deliberate decision not to render aid to Mr. Dunn."
>
> […] "As chief of police, there are times when I wish I could do more. But I'm a firm believer in that good will always win over evil," he added. "It may not come in our lifetime, but there will be justice."

The case blew throughout the American media, making it to talk shows and being shared among celebrities. Another interview by Florida Today (Gallop, 2018) followed the case one year later. They documented that the teenagers suffered death threats, and that there were strange cars in front of the teenager's houses at night.

In this real-life event, the public displayed a great motivation to blame and punish the laughing teenagers for failing to help a drowning man, even though there is no legal obligation of these teenagers to help. In this thesis, I will argue that an important factor that explains these blame ascriptions is the underlying perception of moral character. In both criminal law and in lay people's moral judgments, blame is determined by the presence of two fundamental criteria: a guilty act (i.e., whether or not the agent's action/inaction caused the negative outcome) and a guilty mind (i.e., whether the agent intended and desired the negative outcome). Uhlmann et al. (2014) go further in this discussion and propose that a theory of blame should include people's inherent motivation to identify the morality of individuals. In what they call a person-centered approach to moral judgments, they argue that people blame not only based on perceptions of causality and intentionality (to name a few elements), but they blame because they perceive

immoral individuals. This account seems intuitively fitting for the Jamel Dunn case. Causality and intentionality seem insufficient to explain the huge ostracism against the teenagers, because the young men did not directly cause Dunn's death (e.g., they did not push the man or coerce him to go into the pond), nor directly intended his death. One could argue that even their failure to take action shouldn't stem such strong reactions from the public. After all, people are always at a state of failing to help others, to the extent that harm still exists in the world (for instance, if people are starving in underdeveloped countries, it means that others are failing to help them). What seems to have triggered such strong blame judgments is the fact that the teenagers filmed the man's death and laughed about it. From a legal perspective, filming and laughing has little implication to ascriptions of causality. It is possible that such behaviors suggest a lack of intentionality in helping, yet it is unlikely that an equally unintentional passerby who failed to help would receive death threats for their unintentionality. Rather, a compelling argument for this extreme reaction is that the public perceived the teenagers to be particularly immoral individuals, who deserve to be punished and segregated from society.

In this Ph.D. thesis, I investigated how people form judgments of blame in situations of failure to help like the one cited above and focused on the particular role of moral character inference in explaining such judgments. As explicated, the moral psychological literature has mainly focused on blame for committed acts, whereas it is still an open question whether the same principles that orient blame for proscriptive immorality are applicable for blame for prescriptive immorality. In the following sections, I will provide a comprehensive literature review on the psychological mechanisms of blame.

1.2. Blame in criminal law

An intuitive form of thinking of blame is to address to criminal law. Suppose Bob commits a crime – he shoots a man, killing him. On what basis should people make a moral judgment on Bob's crime? Penal laws naturally differ across cultures, yet there are general principles that guide most criminal justice systems. In the systems that consider a bipartite structure of the crime, liability requires the examination of two aspects of a crime – the act itself (*actus reus*) and the mental state of the criminal when the crime took place (*means rea*). This dichotomy between action and mental states stems from a Cartesian perception of human nature, which divides humanity between body and mind. For an individual to be completely held liable, these two elements should be contemplated (Dressler, 2012).

For complete liability, Bob's conduct, or action (i.e., shooting) should have a causal contribution to the man's death. Taking action (e.g., shooting) is judged more harshly than failing to act so to avoid the situation, that is, omission. However, omissions are still punishable, as long as the context demands the act to be mandatory, that is, when individuals have a legal duty to take action. Fitting to such situations are cases of parents who fail to feed their children, doctors who do not provide appropriate treatment to patients, and policemen who fail to act in order to protect civilians. In the absence of the duty to care, individuals are not legally obligated to help others. As previously mentioned, in the United States, few are the states that adopt the "Good Samaritan law," in which individuals can be punished for failing to help others. This is the reason why, for instance, the teenagers in the Jamel Dunn case could not be prosecuted.

Moreover, for Bob to be completely liable for his crime, Bob's conduct should have been committed with a "guilty" mind, that is, with intentionality, in opposition to negligence. Based on the *means area* criteria, intention refers to crimes that are performed with knowing and desire. On

the other hand, negligence refers to the failure of the duty to care without clear intent, which takes place either consciously or unconsciously (Dressler, 2012).

A drawback from the bipartite system that considers acts and mental states is that it fails to address to justifications and excuses. Imagine that Bob committed the crime of intentionally shooting the man, yet Bob suffers from a mental illness, which clouds his judgment. In such case, the criteria of *actus reus* and *means rea* are both met, yet one could argue about the blameworthiness of Bob. In the tripartite structure of crime, two steps are added to the *actus reus* and *means rea* conditions - the assessment of wrongdoing and of blameworthiness, necessarily in order (Dressler, 2012). According to this framework, for a crime to be liable, it needs to fulfill the criteria of a guilty act, a guilty mind, a wrong act, and a blameworthy act.

What is deemed to be *wrong* is dependent on the social, economic and political context – such are the debatable cases of euthanasia and abortion. Individuals may engage in intentional acts, however the wrongness of it will discern whether this is a moral concern or not. In certain contexts, abortion is accepted when the pregnancy is resulted from rape. The wrongness of intentional actions is also alleviated, for instance, by justifications. A man who intentionally shoots another for self-defense fulfills the guilty mind and guilty act criteria, however his behavior is justified because it was not *wrong*, given the situation.

Finally, blameworthiness refers to whether the offender should be deemed blameworthy for their conduct. Consider again that Bob intentionally shot a man, however Bob suffers from a mental illness. Although there's again a guilty mind, a guilty action, and wrongness to the behavior, Bob can be alleviated from liability because he is excused from blame.

Hence, from a legal perspective, key aspects must be considered in order to assign blame: (1) whether the perpetrator committed an act or failed to act (i.e., omission); (2) the mental state of the perpetrator; and (3) whether there are justifications for the behavior which mitigates its

wrongness; (4) whether there are excuses that exempt the individual of the blameworthiness of their conduct. These factors are consistently explored along the development of the psychological research on blame. In the next section, I will lay out a number of theories and empirical evidence that unravels the psychology behind lay people's judgments of blame and discuss how these theories cast light on the understanding of judgments of blame for failure to help.

1.3. The psychology of blame

Moral psychology has developed much of its understanding of judgments of blame based on the concepts of criminal law. A fundamental concept to understand blame is that of causation.

For a long period, the attribution theories were in fervor among psychologists, and the concept of causation was based on the perception of stable properties of persons or situations. From the perspective of the classic correspondent inference theory (Jones & Davis, 1965), a behavior seems to be caused by one's disposition especially if this behavior does not seem to arise from environmental forces or when it takes place intentionally. In the covariation model by Kelley (1967), a behavior should be attributed to the situation when it consistently evokes the behavior (i.e., consistency), when the behavior occurs distinctively in the given situation but not in other situations (i.e., distinctiveness), and when the situation provokes the behavior among other people (i.e., consensus). A reverse of these patterns would indicate that the behavior is caused by disposition.

During this period, legal philosophers Hart and Honoré (1959) also released their book *Causation in the Law,* which represented a shift of paradigm. They did not explicate causation in the context of the attribution to disposition versus situation. Rather, they explored causation as a fundamental element in predicting blame (Alicke et al., 2015). In their proposition, there are three common sense principles of causation that underlie the attribution of moral and legal responsibility. Specifically, they consider that (1) cause is a necessary condition for the outcome, (2) people infer causes especially for abnormal natural events and human-initiated actions, and (3) this causal inference is weakened when there is interference by another abnormal natural event or human voluntary action, generating a rupture in the chain of causation. Their work was very influential on following research.

For instance, inspired by Hart and Honoré's (1959) causation concept, Fincham and Jaspars (1980) argued that legal rules are likely reflective of common-sense moral judgments. Hence, common law should be useful in revealing the psychological processes of lay attributions of responsibility. In this period, there was a considerate effort to disentangle diverse concepts from the moral research, such as distinguishing causation, responsibility, blame, and punishment. For instance, in their entailment model, Schultz et al. (1981) proposed that causal judgments precede blame, which in turn precedes punishment judgments. They found evidence of such linear paths among 5 to 11 years-old children (Shultz et al., 1985). Fincham and Jaspars (1979) also found suggestive evidence of a precedence of causality over responsibility, and Harvey and Rule (1978) distinguished between blame-praise and responsibility. A great deal of research also applied the concepts of criminal law to examine lay people's moral judgments. The findings suggested that individuals make judgments much like lawyers. For instance, intentional acts were judged more harshly compared to negligent acts (Karlovac & Darley, 1988; Shultz & Wright, 1985); judgments of cause, responsibility and punishment increased for voluntary (vs. less voluntary) actions and omissions (Schleifer et al., 1983; Shultz et al., 1981); and responsibility increased the more the outcome was foreseeable (Shultz et al., 1981). Following this context, robust theories and lines of research built on concepts of criminal law to examine the psychology underlying blame assignment. Due to their relevance for the blame research, in the following sections, I will detail five important theories/lines of research, and I will discuss how these theories contribute to the understanding of blame in the context of failure to help.

Shaver's Theory of Blame (TB). Shaver (1985) proposed an influential theory of blame. He proposed a step-by-step model, much like other stage model theories preceding him - e.g., Piaget's (1965) moral model or Heider's (1958) responsibility model. For Shaver (1985), individuals should search to answer three fundamental questions when assigning blame: "What

caused this event?", "Is anyone responsible for this event?", and "Who is to blame for this event?" The answer to each of these questions are consonant with the ordered components of blame: (1) causation brought about by intentional conduct, (2) responsibility, which is assessed by its dimensions of (a) causation, (b) intentionality, (c) knowledge of the consequences, (d) voluntary choice, (e) the capacity to distinguish right from wrong, and (3) mitigating factors of blame, such as justifications and excuses.

The first assumption to assign blame, as explored by Hart and Honoré (1959), is that the blameworthy negative act or omission must be at least in part *caused* by human action, which should be a necessary condition for the outcome to take place. That is, the answer to "What caused the event?" must include the agent. In the case of multiple possible causes, this model proposes that individuals would make use of the covariation principle (Kelley, 1967) to decide whether the individual represents minimal sufficient cause of the blameworthy event.

The next step is answering the question "Is anyone responsible for this event?" Responsibility in this model is different from causality. Causality is dichotomous and defined either by single cause or multiple causes, and it is "factual." For instance, a tornado can be the cause of the destruction of a city, however a tornado cannot be responsible for it. Responsibility is assigned to humans and refers to a judgment that is established after taking several dimensions into account. For an individual to be responsible for an event, it is necessary that there is involvement of a certain level of personal causation of the individual (first dimension). This causality assessment refers to whether the individual caused the *action/inaction,* rather than caused the *event* (which pertains to the discussed assessment of causation). This is an important distinction in Shaver's (1985) work, because an individual can act without producing any outcome (e.g., crimes of conspiracy), and that is a different situation compared to causing an outcome without acting (e.g., omission). The second dimension relates to the perceived *mental*

state of the agent. Events brought about intentionally are more blameworthy, whereas, in the case of *unintentional* acts, the level of *foreknowledge* of the agent matters (third dimension). No knowledge nullifies moral accountability, whereas the "should have, but didn't know" condition represents minor negligence, and "should know" represents greater responsibility. The fourth dimension is *coercion,* e.g., a threat from a powerful person. In the case of existent coercion, the individual should not be deemed responsible (albeit causal), again nullifying blame, whereas behaviors stemmed from voluntary action are subjected to blame assignment. The fifth dimension of responsibility refers to whether the agent perceives their act to be *wrong,* which would be similar to excuses form the common law and the concept of *capacity*. An individual who cannot recognize the wrongness of their behavior (e.g., an individual with mental illness) should be alleviated of responsibility.

Finally, the last question revolves around "Who is to blame for this event?" Similar to the tripartite system discussed in criminal law, an individual can cause and be responsible for an outcome, however they may still be undeserving of blame. That is when justifications and excuses must be considered. In the existence of plausible justifications and excuses (e.g., self-defense and alleged insanity, as discussed previously), blame should be mitigated. Overall, "when an intentional, voluntary action taken with full knowledge of the consequences and the capacity to understand those consequences is the sole cause of a negative occurrence, the actor is justifiably liable for blame" (p. 172).

Shaver (1985) notes that assigning blame is a complex process and that his model did not intend to provide a rigid explanation to the process of blame. Rather, he specifies each component and its particular order with the purpose of guiding individuals through the understanding of their own blaming process. Hence, Shaver (1985) proposes that the model should be used as a prescriptive framework, or a basic structure, to facilitate consensus. He

predicts that individuals vary in the way they assign causation, blame, and responsibility, and thus intends to provide a predictable pattern through which blame can be understood, allowing individuals to track and identify mistakes in their logic in case of disagreement and correct their blame judgments.

Now, how does Shaver's (1985) work contribute to the understanding of blame for failure to help? First, this model includes the consideration of multiple causes when assessing blame. This is an important distinction, because an event brought about due to failure to help inevitably involves other inferences of causation. For example, in Dunn's case, the teenagers are only one of many possible causes of Dunn's death. One may reason that the man is to blame for his own death, or that any other passerby who might have seen the situation and omitted action should also be deemed blameworthy. Based on Shaver's (1985) theory, the teenagers can still be deemed blameworthy for their failure to help to the extent that they are perceived to be in part responsible for Dunn's death. Shaver's (1985) distinction of causality of events and causality of actions (which is categorized as a dimension of responsibility) becomes an important theoretical contribution in this context, as an individual who fails to help is likely not the sole cause of the event, yet is causal of their own behavior (i.e., whether they acted or not). The same logic applies for intentionality; following the example, whereas the teenagers may not have intentionally caused the unfortunate death, they intentionally chose to omit help. Hence, blame can arise from assessments of responsibility even when the assignment of factual causality and intentionality are blurry.

This account provides an important framework to explain blame judgments that are more complex, like those of omission. Moreover, as explicated, this model assumes that the perceivers who assign blame are bound to errors and have their own motivations to assign blame – which is partly why he established the theory in the first place, so that perceivers could have a system

through which they could evaluate their blaming process and correct for errors. The perception that people make "mistakes" in judgments suggest that there is a normative way through which individuals should blame, based on evidence and rational thought. Shaver's (1985) theory is less accommodating of how the perceiver's motivations and reactive processes influence the process of blaming. These points are explored in the following research.

Blame, heuristics and counterfactual thinking. Concomitant to Shaver's (1985) work was the research on heuristics. Kahneman and Tversky (1982) reported that, under uncertainty, people demonstrate cognitive biases by relying on heuristics. For instance, people try to understand events by reconstructing them in a mental simulation, so that one can test different outputs and assess the extent to which a different outcome could have been produced (i.e., simulation heuristic). In their famous study, Kahneman and Tversky (1982) showed a vignette of Mr. Jones, who got himself killed in a car accident after either taking an unusual route or by leaving work early. When prompted to answer "if only" questions to Mr. Jones' situation, participants tended to undo abnormal events, restoring normality, instead of introducing abnormality. In respect to omission, they also found that participants showed greater regret when outcomes stemmed from actions (e.g., out of two stocks, buying the less profitable one) compared to inactions (e.g., failing to buy the more profitable stock).

The consideration of cognitive biases in moral judgments were extended in Kahneman and Miller's (1986) theory of norms. Their main argument is that norms are not precomputed structures. Instead, norms are constructed on the site after the event has occurred, based on the properties of the stimulus and the context in which it was generated. In their words, "The view developed here is that each stimulus selectively recruits its own alternatives […] and is interpreted in a rich context of remembered and constructed representations of what it could have been, might have been, or should have been" (p. 136). Considering individual's tendency to rely

on heuristics, the assessment of normality is unlikely to follow rational, factual probability. Instead, people are sensitive to events that do not match their expectations, such as aversive and surprising events, and construct alternatives of reality especially for these cases, in comparison to normal events.

This rationale stemmed in important studies on counterfactual thinking applied to the context of moral evaluations, such as blame and causation, as similarly proposed by Hart and Honoré (1959). For instance, in a study by Wells and Gavanski (1989), a man who unintentionally led a woman to die by ordering a dish with wine, to which the woman was allergic, was perceived to be more causal of her death when he chose the dish with wine (vs. an option without wine), compared to choosing a dish with wine paired with another option containing wine. That is, increasing the counterfactual possibilities led to increased perceptions of causation. Another example is that a rape victim received greater judgments of causation, blame and responsibility when participants imagined actions the victim could have taken to undo the rape (Branscombe et al., 1996). In a study by Alicke (1992), a fictional agent was described to have missed a concert due to either negative or neutral acts. When asked how the agent could have avoided this outcome, participants cited negative acts (e.g., getting a flat tire after whistling at a woman at the other car; taking a longer road to pick up drugs from a friend) with greater frequency than neutral acts (e.g., getting a flat tire; taking a longer road due to construction work). Individuals blame not only based on rational assessments of causality, responsibility and mitigating factors. They also blame because they compare people's behaviors with a potential other course of behavior.

The role of heuristics seems particularly relevant when considering blame for failure to help. Societies have evolved with the reinforcement of social norms of cooperation, and, in varying degrees, individuals expect others to help each other. One probable reason why the

teenagers of the Dunn case are deemed blameworthy is because it is easier to think of a second reality in which they did help. Interestingly, helping others is not by itself a normative behavior in terms of frequency (that is, in the large spectrum of demands for help in the world, there is likely more omissions than active helping), yet it is a widespread social norm that dictates behavior. In Dunn's case, the public and the police force were shocked by the teenager's course of behavior and demonstrated the motivation to blame perhaps because it is so easy to think of how they could have behaved differently, in accordance to the social norms of cooperation.

Alicke's Culpable Control Model (CCM). In 2000, Alicke proposed a model of blame cemented in the premise that individuals are capable of controlling their actions and desires. In this model, there are three forms of control that matter for blame, which refer to different combinations of links between mental states, behavior, and consequences. The *behavior control* refers to the link between mental states and behavior, that is, whether the action was freely chosen or not, intentional or not. *Causal control* refers to the link between behavior and consequences, that is, the extent to which one's behavior was the cause of the following consequences. Finally, the *outcome control* refers to the link between mind and consequence, and concerns to the extent an individual desired or foresaw the outcome that took place.

In this regard, Alicke's Culpable Control Model (2000) refines Shaver's (1985) observation of different forms of causality and intentionality (e.g., causality and intentionality for acts and for outcomes), and classifies these components of blame as forms of control. In the perspective of the CCM, a person receives greater blame the more they are perceived to have control through either one of these three linkages, and blame is mitigated when there is constraint of their control. Alicke (2000) argues that his model allows for greater gradation of blame. For example, individuals may assign blame for a person who desires an outcome and benefits from its consequence (the outcome control), even when they were not the necessary cause of the event.

Decision-stage theories like Shaver's (1985) would assume that blame should not ensue once there's no principle of causality, however the CCM accommodates for such situations by allocating blame in the context of perceived control.

Whereas decision-stage theories are modeled to promote justice and rational decision making, Alicke (2000) argues that ordinary perceivers are not always motivated by rationality when making judgments of blame. Indeed, its arguable that the blaming process can head to very distinct directions when an individual is motivated by retributive justice (e.g., a sense of vengeance, or "eye for an eye") compared to a restorative justice (e.g., reinsertion of a criminal in society). A distinct characteristic of the CMM is that it includes the individual's personal expectations, emotional reactions and motivations – which could be perceived as errors in judgments by Shaver's model (1985) – as the basis of his model of blame.

Specifically, the CMM assumes that the process of blame takes place automatically as spontaneous, valenced evaluations, which arise automatically in relevant contexts (Bargh & Chartrand, 1999; Fazio et al., 1986). Such spontaneous evaluations refer to attitudinal positive or negative reactions towards personalities, actions and behaviors when a wrongdoing takes place. In this regard, Alicke (2000) proposes that blame occurs first, and then there is mitigation through cognitive processes. In the absence of a strong reactive evaluation, rational aspects (e.g., desire, foresight, foreseeability, causal influence, coercion, justification) drives the blame judgment (Alicke, 2008). However, when the spontaneous evaluations are strong, individuals may engage in a "blame validation" mode, which means that the evidence is skewed to match the initial evaluation, or that the desire to hold someone culpable affects perceptions of the control so to validate blame (Alicke et al., 2011).

A considerable amount of empirical work supports the CCM. For instance, in his Study 1, Alicke (1992) demonstrates that a person who causes an accident is perceived to be a greater

cause when their reason for speeding is to hide cocaine (vs. to hide a gift to his parents). That is, hiding cocaine (vs. gift) elicited stronger negative evaluations of the perpetrator, which contaminated judgment of blame despite the apparent irrelevance of the information. Now, from Shaver's (1985) perspective, this result could be explained by the concept of justification: hiding a gift can be considered a more justifiable reason for speeding than hiding cocaine. The justification argument becomes less likely in this other study by Alicke and Zell (2009). Their results demonstrated that socially unattractive (vs. attractive) actors received greater blame for harmful outcomes. This effect was reduced when extenuating circumstances were presented *before* participants learned about their character, but not *after*, suggesting that the character information contaminated their blame evaluations. In this case, it is not logical to argue that an attractive character has more of a justification to commit harm compared to an unattractive one. Rather, Alicke (2000) argues that the negative evaluations of the unattractive character led participants to a greater motivation to blame this particular agent.

A variety of studies suggest the role of automatic evaluations on blame. For example, individuals that were characterized as alcoholic (vs. diabetic) were perceived as more causal of their own death (Alicke, 1992; Study 4). Having initial negative (vs. positive) motives led agents to be perceived to be a greater cause of an outcome (Alicke, 1992, Study 4). A man who shot an intruder received less blame when the intruder was a criminal (vs. his daughter's boyfriend), and this effect of victim characterization on blame took place due to the nature of being a criminal, rather than the danger that being a criminal represents (Alicke & Davis, 1989). Although abnormal behavior is known to be more blameworthy due to its greater mutability, when the abnormal behavior is good, blame was discounted (Alicke et al., 2011, Study 2). All of these studies are indicative that valenced spontaneous evaluations (elicited, for instance, by the

morality of the perpetrator) influenced following moral judgments such as those of causation and blame.

The CCM brings an important consideration of the motivational aspect of blame. This model accommodates the findings of the literature on heuristics and counterfactual thinking by considering that individuals are prone to biases in their judgments. In the case of failure to help, it seems sensible that judgments of an agent who fails to help are contaminated by impressions that this failure seems to indicate about the agent. In Dunn's case, the information that the teenagers were laughing when they watched the man drown seems to exert a distinct influence on the public's assessment of blame. From Shaver's (1985) perspective, *laughing* should have little influence on blame, as it is unrelated to causality, responsibility and mitigating factors. However, from Alicke's (2000) account, the information that the teenagers were *laughing* evokes an automatic negative evaluation of the teenagers, which motivates the public to blame them, and possibly ascribe, for instance, greater causal role and intentionality to the teenager's actions so to be able to punish them. The reason people feel negatively about the action in the first place, can also be explained by heuristics of how individuals are expected to behave and principles of normativity, for instance.

Malle et al.'s Path Model of Blame (PMB). In a most recent theoretical work, Malle et al. (2014) propose what they call the Path Model of Blame. According to this theoretical framework, blame is a cognitive and a social process. Cognitive, because individuals have developed their psychological apparatus capable of identifying norm violation scenarios, processing information such as intentionality, preventability, causality, to name a few, and generating a blame response. Social, because blame's fundamental purpose is to regulate behavior of individuals of a collective. In that sense, Malle et al. (2014) emphasizes that blame is a process directed to individuals involved in norm violations with the purpose of regulating

behavior. Blame is also a costly process. Blaming is intended to correct behavior and can cost the reputation of the "corrected" person, therefore blaming per se requires warrant.

Because of the important social consequences of blame, indiscriminate blaming is socially penalized, and people expect one to clearly show evidence of why one should be blamed. These *why's* rely on cognitive judgments, which occur in a stepwise manner as individuals evaluate blame. These steps are defined as a *default processing order*, however the processing order may be loosened in more complex situations.

The initial component of the so-called PMB is detecting a deviating event/outcome, assessing that an agent caused it, and deciding whether this was brought about intentionally (necessarily in order). Once this decision is made, the process of blame divides into two different pathways. If the act is evaluated as *intentional*, one should consider the agent's subjective reasoning, the causal background of those reasons (such as personality, culture and context) and enabling factors that allowed the agent to fulfill that behavior. On the other hand, if the behavior is believed to be brought about *unintentionally*, then the perceiver considers whether the agent should have prevented the event/outcome (i.e., obligation) and whether the agent could prevent the event/outcome (i.e., capacity). A capacity limitation may refer to physical strength of time constriction, for instance. Figure 1 illustrates their model.

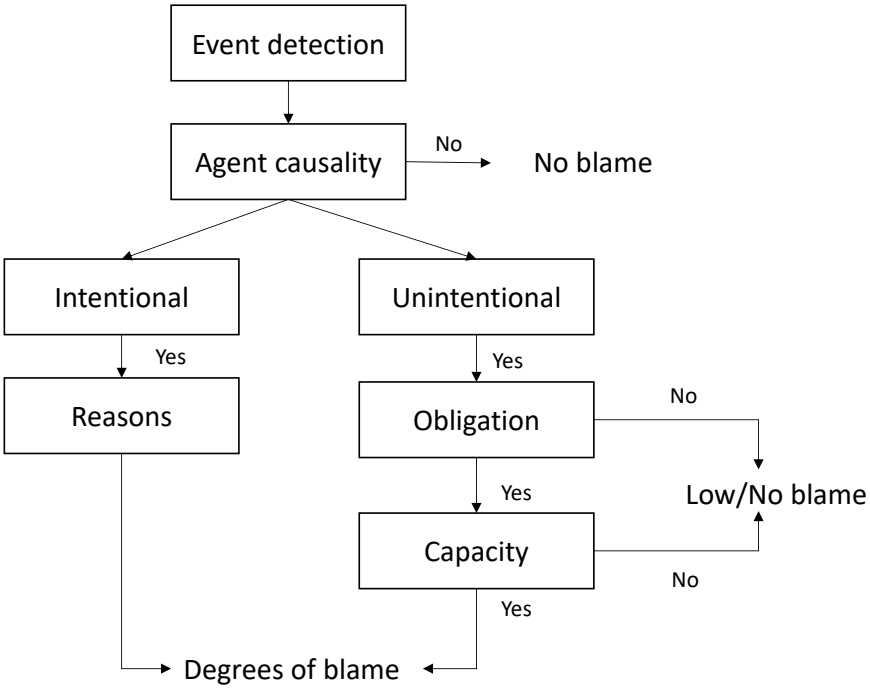

Figure 1. A visual representation of the Path Model of Blame by Malle et al. (2014).

There is convincing empirical evidence for this theory. To test for the stepwise nature of blame assignment, Guglielmo (2012) and Guglielmo and Malle (2014) had participants go through several norm-violating events and investigated the types of information the participants would seek to make blame judgments of the events, either in an open ended version (in which participants openly asked information of the events to supply their understanding of the events) or in a guided version (in which they asked participants whether they wanted information on causality, intentionality, reasons, obligation, and justification). In both approaches, the results showed that participants sought for information in the hierarchical order proposed by the PMB.

In a more recent paper with six studies, Monroe and Malle (2018) found that people consistently update blame judgments by taking into consideration elements predicted by the PMB. Importantly, they also found evidence that these updates take place symmetrically. For

instance, people exacerbate blame for an intentional act with bad reasons to a similar extent that they mitigate blame when the reason is good, and they exacerbate blame for an unintentional and preventable act to a similar extent that they mitigate blame when the act is unintentional and unpreventable. With this finding, Malle et al. (2014) go against the motivational approach such as Alicke's (2000). If people were to show a motivation to blame, they should be resistant to mitigate blame after their initial judgment. The symmetry result was found for different samples (i.e., students, Internet, and community) and persisted when participants were subjected to a cognitive load task, suggesting that people's precision and flexibility in making moral judgments is likely a default pattern of processing blame. Curiously, in their Study 6, Malle et al., (2014) found that the symmetric judgments disappeared when people judged outgroups (vs. ingroups). Individuals blamed outgroup members to a greater extent than ingroup ones and were more reluctant to diminish blame after learning of mitigating factors (i.e., one's good reason for intentional acts and unpreventability for unintentional acts).

The PMB is a refined theory which proposes that individuals assign blame with the purpose of regulating behavior, and they do so carefully, as blame requires warrant. A fundamental distinction of this theory with others is the prediction of two different paths of processing information based on intentionality. It organizes the process of blame assignment based on hierarchical evaluations of causality, intentionality, followed by either reasons or obligation and capacity, yet it also stipulates that people may deviate from this stepwise process when judging morally complex information. Based on this model, it's possible to predict that judgments of blame for failure to help are based on perceptions of causation and intentionality and should be mitigated by perceptions of justification, obligation and capacity.

Moral character and blame. Finally, Uhlmann et al. (2015) propose what they call a person-based approach to moral judgments. They argue that stage models account such as

Shaver's (1985) TB and Malle et al.'s (2014) PMB miss an important factor of human motivation – the motivation to evaluate the moral character of others. When making moral judgments such as that of blame, individuals do not ask themselves only whether the act performed was good or bad, right or wrong, but they also fundamentally evaluate whether the agent who performed the act is a good or a bad person. In this sense, humans make judgments not as consequentialists interested in the best outcome, but as intuitive virtual theorists, constantly searching for the moral character underlying people's behaviors.

Their first argument to defend the "people as intuitive virtual theorist" framework is that judgments of moral acts are not necessarily correspondent to the judgments of the moral person behind the act. For instance, in Uhlmann et al.'s (2013) study, a hospital administrator who spent $2 million on hospital equipment and saved 500 lives (instead of funding a single life-saving operation for a little boy) was perceived as more deficient in moral character, despite making the more pragmatic and praiseworthy decision. Other studies showed that participants inferred greater immoral character of an agent who performed racial slur versus physical assault, even though the latter was judged as a greater immoral act (Uhlmann et al., 2013). Likewise, hitting a cat was more informative of one's immoral character than hitting one's girlfriend, even though hitting a girlfriend was perceived as the more immoral act (Tannenbaum et al., 2011). These findings suggest a dissociation between moral judgments of acts and persons.

Now, if people are motivated to make moral judgments on the basis of character, then acts that are performed by worse characters should receive greater retaliation. Supporting evidence to this assumption shows that people judged harmless acts more harshly when it was highly diagnostic of immoral character (Tannenbaum et al., 2011), and that an immoral (vs. moral) agent received greater blame for committing harm (Nadler, 2012). This account also explains the findings of harsher judgments towards immoral agents (Alicke, 1992; Alicke & Davis, 1989)

from a moral-character approach. Different from Alicke (2000), Uhlmann et al. (2015) argues that exacerbated blame for immoral individuals does not constitute a bias. In their view, character evaluation is an essential aspect of the moral system and is taken as another aspect that individuals consider when making judgments, rather than a bias. According to this person-centered approach to moral judgments, blame serves the function of removing "bad people" from society (Pizarro &Tannenbaum, 2011; Uhlmann et al., 2015).

The person-based approach to moral judgments offers a novel element in understanding judgments of blame for failure to help. Based on this account, individuals are blamed not only due to their actions, but also on the basis of what their act reveal about their moral character. This account seems to offer an intuitive explanation to the motivation to blame the laughing teenagers who failed to save Jamel Dunn. People are blaming the teenagers not only because of their causal role or because they could have prevented the outcome, but also because they seem to be particularly immoral characters.

Implications for studies on blame for failure to help. In this section, I presented five important theoretical frameworks in understanding the psychology of blame – Shaver's (1985) Theory of Blame; the studies on heuristics and counterfactual thinking; Alicke's (2000) Culpable Control Model; Malle et al's (2014) Path Model of Blame; and the person-based approach of moral judgments. Table 2 shows a summary of these theoretical frameworks. The most notable convergence among the theories in respect of the fundamental elements of blame (especially the TB, the CCM and the PMB) is the consideration of intentionality and causation. Both elements occupy the primary positions in the process of designating blame in the TB and the PMB and constitute elements of control in the CCM perspective. Much like in criminal law, lay people incorporate the criteria of *actus reus* and *means rea* in everyday accounts of blame judgments.

Table 2. Summary of theoretical frameworks on blame

Model and Author	Keywords	Proposition
Theory of Blame (TB): Shaver (1985)	Theory of blame	Step model of blame. Blame determined by (1) causation, (2) responsibility, composed by causation, intentionality, knowledge, coercion, capacity to acknowledge wrongness, and (3) justification and excuses
Heuristic Theories: Kahneman & Tversky (1982); Kahneman & Miller (1986)	Heuristics; counterfactual reasoning	Blame is influenced by individual's expectations and heuristics. Blame is influenced by how easy it is to undo the negative outcome through simulation.
Culpable Control Model (CCM): Alicke (2000)	Personal control; spontaneous valenced evaluations; blame validation	Blame is determined based on perceptions of personal control. Blame is influenced by spontaneous valenced evaluations. People engage in blame validation.
Path Model of Blame (PMB): Malle et al. (2014)	Step model of blame; two pathways based on intentionality	Once negative event is identified and causal role established, blame process unfolds depending on whether the act was intentional (mitigation through reasons) or unintentional (mitigation through obligation and capacity)
Person-Centered Approach to Moral Judgments Uhlmann et al. (2015)	Moral character	People blame based on assessments of moral character

This thesis is focused on how individuals blame failure to help. Because of the centrality of the roles of action-inaction and mental states on blame, I will discuss these two elements in greater detail in the next sections, whilst associating it to the particular context of blame for failure to help. Individuals likely blame a person who fails to help based on their behavior, which is closely related to ascriptions of causation (e.g., did they fail to help by taking action or by omitting action?) and based on evaluations of their mental states when they fail to help (e.g., did they fail intentionally or unintentionally? Did they bear positive or negative mental states when they failed to help?). However, blaming failure to help represents a more complex moral judgment because it typically entails multiple causation. From a strictly consequentialist

perspective, a person who fails to help another should not be recipient of any blame, because the negative consequence takes place regardless of their action. However, it is possible that blame still ensues at the perception of "sufficient" blame, as proposed by Shaver (1985). The three models predict that blame may ensue especially due to perception of preventability (in the CCM's terms, the link between behavior-outcome controls), which is reflective of a counterfactual thinking process. The PMB's account includes the evaluations of obligation and justifications as mitigating factors.

One open question is whether individuals reach these judgments in an unbiased manner or not. Based on the literature of heuristics, it is possible that people may, for instance, overestimate a person's capacity to prevent an outcome to the extent that it seems correspondent to heuristics of moral norms. For instance, in Dunn's case, people may expect that the teenagers should have helped based on shared, moral principles prevalent in society (even though, logically speaking, the frequency of people omitting help is superior to those offering help), and hence may overestimate their capacity of preventing the man's death, or underestimate their justifications. This possibility is also explored in the current thesis.

Finally, based on the person-based approach to moral judgments, it is also possible that people blame failure to help to the extent that such failure seems indicative of immoral character. Traditional attributional theories have long posited that people use both information of behavior (Jones & Davis, 1965) and intentionality (Heider, 1958) to infer disposition. Therefore, it is theoretically meaningful to investigate whether the effects of mental states and actions on blame can be partly explained by inferences of character. This rationale will also be further explored in the next sections.

1.4. Action-inaction, moral character, and blame for failure to help: Theoretical interpretations

A crucial information for assessing blame is to evaluate whether the negative event was brought about by an agent's action (i.e., commission) or inaction (i.e., omission). The moral literature has shown significant implications to moral judgments when contrasting between action and inaction at the empirical level. In this section, I will first present the relevant research on this domain. Then, I will provide a framework based on the previous models of blame to conceptually explain this action-inaction effect on blame.

From a consequentialist viewpoint, the distinction of commission and omission is irrelevant, to the extent that they equate in terms of outcomes. Therefore, Spranca et al. (1990) first refers to the preference for inaction (vs. action) when there's a negative outcome as a bias – the *omission bias* – because there is an assumption that the consequentialist view should be the logical perspective. In six studies, they found evidence for this inaction preference. For instance, in one of their studies, they asked participants to judge how moral John's behavior is when, with the purpose of making his competition sick, he either recommends his competition (Ivan) to have a house dressing to which Ivan is allergic to (i.e., action) or does not stop Ivan from having the said dressing (i.e., omission). Participants evaluated the behavior as worse when John committed the negative act (vs. inaction). The omission bias took place even for a within-subjects design, and when they actively asked participants to compare the situations in written form and paid them for their time. The bias was also present after having controlled for intentions, outcomes, knowledge, and effort.

Ritov and Baron (1990) followed Spranca et al.'s (1990) work. They found that people prefer to omit vaccination even when the outcomes of the inaction are worse than that of

vaccinating. People feel more responsible for comissions (vs. omissions) and missing out information increases people's reluctance to vaccinate. They also showed evidence that the omission bias takes place even after controlling for the status quo bias, and that it cannot be explained by perceived harm, as the omission bias persists even when the omission is more harmful than commission (Ritov & Baron, 1994). Later studies followed showing the bias for judgments of blame, wrongness and punishment (Cushman, 2008; Cushman et al., 2006) in several moral domains (DeScioli et al., 2012).

There are three important explanations to the omission bias, namely (1) perception of causality, (2) transparency, and (3) heuristics. As previously mentioned, commissions are perceived to be more causal than omissions (Kordes-de-Vaal, 1996; Ritov & Baron, 1990). Whereas acts are perceived to be the sole cause of an outcome, inactions allow for multiple inferences of causation. Negative outcomes brought about by inaction (vs. action) tend to happen naturally (e.g., a drowning man would naturally die anyways even if the omitter had not been present), to take place indirectly and to allow for defusal of responsibility (see Baron & Ritov, 2009).

More recently, DeScioli et al. (2011) argue that omissions tend to be judged more leniently because the processes underlying its occurrence are more unclear. Just as causality, it is also harder to infer the intentions underlying omissions (Kordes-de-Vaal, 1996; DeScioli, 2011), as well as the potential reasons for its occurrence. However, when there is transparency of the decision-making process, omissions should be as condemnable as commissions. In their study, they contrasted an "opaque" omission condition, in which the decision-making process is unclear (i.e., an agent chooses not to push a button that could save another individual from dying), with a "transparent" omission condition (i.e., the agent chooses to push a button that has no consequence in helping the individual instead of one that would). The results showed the agent who

"transparently" chose to omit help was punished more than an agent who omitted help in the "opaque" condition. Moreover, the transparent omission was judged as harshly as pushing a button that directly caused the victim's death (i.e., commission). This effect was replicated even when controlling for perceived intentionality and causality. To the extent that the choice to omit help is transparent to the public eye, Descioli et al. (2011) argue that omissions can be as blameworthy as commissions.

Finally, according to the person-centered approach to moral judgments, blame is assigned based on the perceived immorality of the agent. Uhlmann et al. (2015) argues that this approach may explain the omission bias to the extent that individuals reason that "it takes a particularly bad person to commit negative acts", whereas omission should be less diagnostic of character. People infer disposition based on behaviors (Jones & Davis, 1965), and some behaviors are more telling of character based on heuristics. For example, people tend to hold protected values (i.e., values perceived as sacred and unavailable for trade-off) to a greater extent for commissions (e.g., "do not kill," or "do not hurt others") than omissions (e.g., "do not allow others to kill" and "do not allow others to be hurt"). Omissions are also more prevalent and expected compared to commissions of negative acts (Baron & Spranca, 1997). As shown in the social perception literature, acting out on negative behavior is particularly diagnostic of character (Mende-Siedlecki et al., 2013).

It is relevant to notice that the literature on omission bias has primarily investigated omission (i.e., a failure to act so to prevent a negative outcome), by contrasting it with commission (i.e., in which one actively commits a negative outcome). This comparison allows for an interpretation of how inaction is judged differently from a committed negative act, with a focus on proscriptive immorality (i.e., relative to behaviors that one should not commit). However, this approach is not informative of how omissions per se are blamed. In other words,

whereas it is useful to know how "killing" is more blameworthy than "letting one kill", it is still unclear how "letting one kill" is blameworthy at all. This thesis is focused on blame in the prescriptive immorality domain, which refers to failure to help. A consideration of an inaction-action effect in this scenario should entail a comparison between an individual who takes action to help (i.e., commission) versus one who omits from helping (i.e., omission). This refinement in the operational definition of omission is an important feature explored in Study 1 of this thesis, allowing to capture the effect of omission by itself on blame.

Specifically, when comparing an individual who fails to help after taking action (i.e., failed attempt of helping) with an individual who fails to help after not taking any action (i.e., omitted help), it seems intuitive that taking action presents as a preferable decision. In this sense, the action-inaction effect for prescriptive morality would unfold in an *action bias* (i.e., a preference for action versus inaction in the face of negative events), instead of an omission bias. This seems plausible for all explanations of the omission bias. First, a person who takes action to help and fail provides more evidence of the lack of their causal role in the negative event compared to one who omits help. Second, taking action to help is a more transparent situation than omitting help, in the sense that it leaves little room for questions of causality, intentionality, and preventability, and justification, which are relevant factors predicted by the theories of blame. Finally, as Baron and Ritov (2004) discuss, just as the omission bias, *action bias* arises in situations in which the heuristics demand for action (e.g., "Don't just sit there, do something!"). That seems to be the case for situations in which one is called to help others, as people are socially reinforced to take action to help rather than omit help. Failure to take action may seem as particularly diagnostic of an immoral character ("only a bad person would not try to help"), which may explain greater blame for inaction (vs. action).

Hence, Study 1 of this thesis explores the possibility of an *action bias* in blame for failure to help, that is, the prediction that people will show a preference for individuals who take action to help before failing to help compared to an individual who does not take action to help, hence failing to help. This study represents an application of the robust action-inaction effect on blame for committed negative acts (i.e., proscriptive morality) to the context of blame for failure to commit positive acts (i.e., prescriptive morality).

Summary. There is ample evidence in literature of an omission bias in the context of proscriptive immorality. This bias has three important explanations: (1) commissions are perceived as more causal than omissions; (2) commissions are more transparent in terms of underlying factors such as intentionality, preventability and justification compared to omissions, which are more ambiguous; and (3) commissions are more diagnostic of immoral character compared to omissions. However, the action-inaction effect has been mainly explored in the context of proscriptive immorality, and less is known about its effect on moral judgments in situations of prescriptive immorality, that is, failure to help. It is likely that the effect action-inaction on blame for failure to help takes place with a preference for actions (vs. inactions). That is, omitting to help is likely more blameworthy due to perceptions of causality, lack of transparency and inferences of immoral character. This hypothesis is the focus of Study 1 of this thesis.

1.5. Mental states, moral character, and blame for failure to help: Theoretical interpretations

As explicated at the beginning of Chapter 1, another criterion for judging a crime is to judge the perpetrator's guilty mind. In this section, I will provide empirical evidence of the effect of mental states on blame, and once more discuss potential explanations to this effect based on the theories provided in Section 1.3.

Human beings have a tendency to assume that other individuals have a mind – that they have wants, thoughts and beliefs of their own, and that such mental states can be useful in predicting one's behavior (Premack & Woodruff, 1978). The ability to perform mindreading becomes particularly sophisticated among children from 2 to 7 years old, related to the development of language and executive functions (see Apperly, 2011). Classic research on the moral development shows that children show increasing capacity of weighting intentions (vs. consequences) when making moral judgments (Constanzo et al., 1973; Piaget, 1965).

From an early age, humans are sensitive to information about other people's positive and negative intentions. For instance, Hamlin (2013) exposed infants of 5 and 8 months-old to a puppet show in which the puppet either helped or hindered a third party in their goal of opening a box, varying in whether they succeeded or not and whether the outcome was consistent with their intentions. Overall, 8 months-old showed preference for the helper puppet regardless of the outcome of their intentions and did not prefer better outcomes when the intention was constant. In another study, Hamlin et al. (2013) found that 10-months-old infants showed preference for puppets who helped a third puppet reach a toy it preferred (vs. a puppet who helped the third puppet reach a toy that it didn't prefer). Infants are sensitive to the positive and negative mental states of puppets and favors puppets who are helpful. Overall, these findings suggest that mental

state information affects judgments of morality early on in development. Some theories on morality go to the extent of affirming that mental states are the basis of morality (Gray et al., 2012).

The information of the perpetrator's mental state becomes even more fundamental for moral judgments in adults. Ample evidence shows that greater intentionality of committed immoral act is associated with harsher moral judgments. Intentional (vs. unintentional) acts are perceived to be more harmful, even when they are not (Ames & Fiske, 2013). A man is considered as more of a cause of their partner's death when he poisons the partner intentionally (vs. accidentally) (Alicke, 1992, Study 4). Rapist that admit intent receive greater blame and punishment for their crime (Kleinke et al., 1992), Furthermore, such effect of mental states takes place even in the absence of tangible consequences; people judge acts that are intended as more blameworthy than unintended ones, even when none of the acts produced any outcome (Cushman, 2008). Like lawyers, lay people also judge acts performed with evil intentions as morally worse, regardless of the related outcome. As introduced in the previous section, intentionality is a core concept in judgments of blame.

The literature on the effect of mental states on blame for negative acts is extensive. However, there is little empirical data on how mental states play a role in blame for failure to help. The literature has mainly focused on blame for proscriptive immoralities (i.e., blame for acts that should have been inhibited) instead of prescriptive immoralities (e.g., blame for failure to help), hence studies that have investigated individuals' mental states when helping have typically assessed its effect on praise, and in a context of comparison with blame. For instance, Pizarro et al. (2003) observed that individuals who help impulsively (vs. deliberately) were praised similarly, whereas individuals who harmed impulsively (vs. deliberately) were less blamed. Malle and Bennett (2002) showed that positive intentions predicted praise to a lesser

extent compared to how negative intentions predicted greater blame. Hence, the present work focuses on the effect of both positive and negative mental states on blame for failure to help.

To investigate the effect of mental states on blame for failure to help, it is important to cite Malle and Knobe's (1997) distinction between intentionality and intentions. The authors propose that intentions entail one's *desires* for an outcome and the *belief* that a certain action will lead to a particular outcome. Intentions are designated to an agent (i.e., a person has intentions). On the other hand, an act is only considered to be *intentional* if the actor has the *skill* to perform the action and *awareness* of fulfilling the intention. Intentionality refers to the act (i.e., whether an act takes place intentionally or unintentionally). Although Malle and Knobe (1997) make a careful distinction between intentions and intentionality, for the lay judger many times these two concepts are conflated. For instance, in their own study (Study 3), participants explicitly defined what an act performed intentionally is, and their answers were coded into the categories of desire, belief, intentions, and awareness (the category of skill was only implemented in Study 4). Fifty percent of the responses were categorized under intentions, showing that people don't distinguish all five aspects of intentionality at once when freely defining it on a task, and that people may conflate their definitions of intentions and intentionality in a free form of response. This point is considered in their later work (Malle et al., 2014), "Even though people are highly sensitive to these five components […] they do not deliberate about the components each time they judge whether the behavior is intentional. Instead, they quickly recognize intentionality in everyday situations" (p.154).

Referring once again to Dunn's case, the teenagers failed to help the man and endorsed in questionable behaviors (laughing and filming). The fact that they laughed and filmed suggest that they had little desire to help, and that they knew the situation, that is, that they had no intentions of helping. It is more complicated to infer whether they failed to help intentionally or

not. Intentionally failing to help would presuppose that they used their skills and awareness for a deliberate decision to not help, which sounds unnatural for the case. Intentionality refers to acts, yet failure to help typically takes place by inaction. Considering this complication and that intentions is a more nuclear concept for mental states inferences (e.g., intentions is one component of intentionality), in this thesis, I focused specifically on the effect of intentions on blame, rather than intentionality. This decision is in line with criminal law, in which the mere presence of intentions is enough to constitute a guilty mind.

How does one's bare intentions affect blame judgments for an individual who fails to help? The person-centered approach to moral judgments suggests that people blame based on inferences of moral character. To the extent that intentions are informative of disposition (Heider, 1958), a possibility is that negative intentions increase judgments of blame because they signal that the agent is a bad person who should be punished. Another possibility is that positive intentions may lead to perceptions of moral character of the agent, which may in turn alleviate blame. Moreover, the literature on the blame-praise asymmetry suggests that the effect of intentions on blame should be stronger for negative (vs. positive) cases.

In Study 2 of this thesis, I examined the first proposed path – that is, whether bare intentions are indicative of moral character. For a more extensive understanding, I investigated the potential psychological mechanisms for this effect and whether the effect of negative intentions is stronger than that of positive intentions. Evidence in literature suggests that intentions alone affect inferences of character. For instance, Cohen and Rozin (2001) found that inner states of disliking one's own parents or fantasizing about an affair (unfollowed by actual behavior) were informative enough to make a person a worse character, at least among Protestants. Other studies revealed that agents were judged as immoral even when they were not the cause of harm, to the extent that they experienced pleasure at others' harm (Gromet et al.,

2016), and desired the harm to happen for an unrelated reason, i.e., “wicked desires” (Inbar et al., 2012). These studies indicate that an agent’s mind is by itself sufficient for people to judge this agent to be immoral. However, less is known about whether people are deemed to be good simply on the basis of having positive intentions. Based on Kelley’s (1967) reasoning, positive intentions may be less diagnostic of character to the extent that they are supposed to be consistent and consensual in society, and that people seem to expect others to have positive intentions (Pizarro et al., 2003), which may diminish attributions of disposition. The role of informativeness of intentions is explored in Study 2 as a potential mediator to the effect of intentions on blame.

In Study 3 of this thesis, I examined the role of moral character inferences in the effect of positive and negative intentions on blame in scenarios in which individuals failed to keep a promise to help. As mentioned, this hypothesis was based on the person-based approach to moral judgments (Uhlmann et al., 2014).

1.6. Overview of empirical findings

Studies and theories on blame have mainly focused on how blame is assigned for proscriptive immorality, that is, for actions that one should not do. Less is known about how individuals reach judgments of blame for prescriptive morality, that is, for failing to do what one should do (e.g., failing to help). This thesis is focused on the psychological processes underlying blame for failure to help. There are two fundamental information that are considered when one assigns blame: (1) whether the negative outcome was brought about by the individual by an action or an inaction, and (2) the mental state of the agent when they incited the negative event. In the following studies, I have explored each of these elements separately. Based on the person-based approach to moral judgments, I also investigated whether moral character is an important explanation to the effect of these elements on blame. Exploring the specific processes underlying blame for failure to help and testing the moral character explanation for the effect of both action and mental states on blame are important theoretical contributions of this work.

Below, I present an overview of the studies that compose this thesis.

Study 1. Action-inaction effect on blame for failure to help. In a set of three studies, we examined the psychological processes underlying blame for to omitting help by contrasting blame judgments for agents who fail to help either by taking action and failing (i.e., action condition) or by omitting help (i.e., inaction condition). We demonstrate that people blame an individual who omitted help to a greater extent than one who attempted and failed to help. In three studies, inferences of moral character partially explained the omission effect on blame even after taking into account elements of causality, intentionality, capacity, and justification. This study is reported in the following manuscript:

Hirozawa. P. Y. & Karasawa, M. (under review). Moral character evaluations underlie blame for choosing to omit help.

Study 2. Mental states and assignment of moral character. The literature on person perception suggests that negative actions are more diagnostic of character than positive actions (Reeder & Spores, 1983; Riskey & Birnbaum, 1974; Skowronski & Carlston, 1987). In this set of two studies, I investigated whether negative intentions are also more diagnostic of character than positive intentions. The results showed that negative intentions are more predictive of immoral character than positive intentions are predictive of moral character. Two psychological mechanisms were examined, namely moral emotions and informativeness. Mediational analyses revealed that emotions mediated the effect of intentions on character inference for both positive and negative intentions, however diagnosticity was a significant mediator only for negative intentions. That is, both positive and negative intentions are diagnostic of character, however negative intentions seem to be particularly telling of immoral character due to its informativeness. This study was published in the form of the following article:

Hirozawa, P. Y., Karasawa, M., & Matsuo, A. (2020). Intention matters to make you (im)moral: Positive-negative asymmetry in moral character evaluations. *Journal of Social Psychology, 160*(4), 401-415.

Study 3. Mental states effect on blame for failure to help. In this study, we tested whether the valence of mental states (i.e., desires) affects judgments of blame for failure to help due to inferences of moral character. Specifically, six scenarios described an agent who made a promise yet failed to fulfill it, generating a negative outcome for the promised person. We manipulated whether the agents held positive, neutral or negative desires to help. In all cases, the failure to help occurred due to forgetfulness. Results revealed that participants assigned greater blame to an individual with negative (vs. positive) desires. One again, moral character explained

this effect on blame, as well as perceptions of wrongness. This study is reported in the following article:

Hirozawa. P. Y. & Karasawa, M. (2020). Negative desires make failure to help more blameworthy: The role of wrongness and moral character evaluations. *Journal of Human Environmental Studies, 18*(2), 119-126.

Summary. The present thesis explores the psychological processes underlying blame for prescriptive immorality, that is, failure to help. A central investigation was to test whether inferences of moral character explains how individuals blame failure to help. Study 1 demonstrates that an agent who omitted help (vs. attempting but failing to help) received greater blame for the consequences of their failure to help in part because they were perceived to be more immoral characters. Empirical evidence shows that negative actions are more diagnostic of character than positive actions (Reeder & Spores, 1983; Riskey & Birnbaum, 1974; Skowronski & Carlston, 1987). In Study 2, we examined whether negative intentions are also more diagnostic of character than positive intentions. We found that negative intentions are particularly indicative of immoral character due to emotional reactions and due to perceptions based on consistency and consensus of the intentions. Finally, Study 3 investigated whether people blame an individual with negative (vs. positive) mental states to a greater extent due to inferences of moral character. The results showed consistency with this hypothesis. Overall, these three sets of studies show evidence that an important mechanism through which people blame failure to help is by assessing the moral character of the omitter.

Chapter 2 - Study 1. Action-inaction effect on blame for failure to help

The present study examines how people are deemed blameworthy for omitting help by contrasting cases in which an agent fails to help either by taking action prior the failure (i.e., action) or by omitting help (i.e, inaction). As discussed in Chapter 1, from a legal perspective, it is rarely the case that people are blamed for failing to help. Only a few states in the United States have adopted the "Good Samaritan" law, and generally individuals are not obligated to help others, unless there they have a legal duty to do so (Dressler, 2015; Keiler et al., 2017). However, people assign moral obligation to help even in the absence of a legal obligation (Buckwalter & Turri, 2015), suggesting that omitting help is worthy of blame. When the process of deciding to omit help is transparent to the public, individuals blame omission to the same extent as commissions (DeScioli et al., 2011).

In this study, we examined the psychological mechanisms underlying blame for deliberate decisions of omitting help. Preceding research (Cushman et al., 2006; Ritov & Baron, 1990; Spranca et al., 1991) has typically investigated omission, that is, a failure to act so to prevent a negative outcome, by contrasting it with commission, in which one actively commits a negative outcome. However, when blaming omission (e.g., not acting to save one in need), people are likely comparing it with a scenario in which the action occurs (e.g., trying to save a person in need), instead of a commission scenario (e.g., killing). Hence, to capture the effect of omission by itself on blame, we compared cases of failure to help by omission (i.e., inaction) with cases of attempted but failed help (i.e., action). This refinement in the operational definition of omission is an important feature of this study.

Importantly, in both inaction and action conditions, the agent's failure to help is followed by a negative outcome. From a utilitarian perspective (Greene et al., 2008; Mill, 1998), taking

action or not should be irrelevant for assigning blame, as both decisions result in the same negative outcome. This conclusion can as well be predicted when reasoning from the legal perspective. However, two important lines of moral research allow us to predict an effect of omission on blame.

According to the framework in which people make moral judgments as "intuitive lawyers" (see Alicke et al., 2015; Fincham & Jaspars, 1980; Hamilton, 1980), people assign blame based on several legal principles, such as the degree to which the agent *caused* and *intended* the outcome, had an *obligation* to help, was *capable* of preventing the outcome, and had a good *justification* to behave the way they did (Cushman & Young, 2011; Hamlin, 2013; Lagnado & Channon, 2008; see Malle et al., 2014). Allowing a negative outcome to happen (vs. actively bringing it to fulfillment) tends to be perceived as less causal of the following negative outcome (Baron & Ritov, 2004; Bostyn & Roets, 2016; Spranca et al., 1991). Likewise, omitting help (vs. helping) may be perceived as less causal of a potential positive outcome (e.g., safety of the person in need) and subsequently increase perceived causality for negative outcomes that follow failure to help (e.g., harm to the person in need). Inactions also tend to be assigned lesser intentionality (Hayashi, 2015; Kordes-de-Vaal, 1996), and effort (see Albarracin et al., 2019). Moreover, people tend to assume that others have a moral obligation to help (Buckwalter & Turri, 2015), so not taking action to help should require greater justification. Thus, based on these findings and rationale, we considered these five components - causality, intentions, obligation, capacity and justification - to be important potential mediators of the predicted omission effect on blame. As referred in Chapter 1, these are important components in predicting blame (Malle et al., 2014; Shaver, 1985).

The second line of research frames moral judgers as "intuitive virtue theorists" who assign blame by evaluating the moral character underlying people's actions. According to this

person-centered approach to moral judgments, blame serves the function of removing "bad people" from society (Pizarro &Tannenbaum, 2011; Uhlmann et al., 2015). Supporting evidence shows that people judged harmless acts more harshly when it was highly diagnostic of immoral character (Tannenbaum et al., 2011), and that an immoral (vs. moral) agent received greater blame for committing harm (Nadler, 2012). Uhlmann et al. (2015) argue that inferences of moral character may underlie the tendency to blame commissions to a greater extent than omissions, as it takes "worse people" to actively harm others (vs. allowing harm to happen). We extended this rationale to the context of blame for omitting help. Inactions are perceived less positively than actions (Albarracin et al., 2019), and individuals are morally expected to help others (Janoff-Bulman et al., 2009). It is plausible that choosing to omit help should be indicative of greater immoral character than attempting to help and failing. Hence, we also tested whether increased perceptions of immoral character explain the effect of omission on blame. This step is a highlight of this study. To our knowledge, no previous research has investigated the omission effect from a person-based approach while considering the cognitive aspects of causality, intentions, justification, capacity and obligation.

Study 1.1

Method

Participants

Japanese university students (N = 111) participated in the study for course credit (69 males, M_{age} = 19.97, *SD* = 0.88). We determined the sample size based on previous analyses using G*Power for repeated measures Analyses of Variance (ANOVAs), within-between interaction (η_p^2 = 0.02, 80% power, four groups, three measurements, correlation of 0.60 among repeated measures, and nonsphericity correction equal to 1), indicating the minimal sample size of 96. There was no exclusion of participants.

Materials, design, and procedures

We prepared two sets of three scenarios in which an agent failed to help, followed by a negative outcome for another individual. The first set depicted direct negative outcomes to individuals (e.g., failing to save a drowning swimmer, who came to die), whereas the second set represented indirect harm through material damage (e.g., failing to summon help for a house on fire). As previously described, we also manipulated whether the agent failed to help by taking action or not. For example, the agent saw a drowning man and judged that there was a possibility that he could save him. The agent proceeded by either doing nothing, believing there was nothing he could do (i.e., inaction condition) or by swimming to the man but failing to reach him in time (i.e., action condition). In both cases, the drowning man dies (for full description, see Appendix 1.1). Each participant was randomly assigned to read one of the three scenarios (in randomized order) pertaining to either the direct harm or the indirect harm condition in a 2 (Omission: action, inaction) x 2 (Harm: direct, indirect) x 3 (Scenarios) mixed design, with Scenarios as within-subject. After reading each scenario, participants rated the following items on 6-point scales.

Blame. Participants rated how much they blamed the agent for the negative outcome (*1 = no blame at all, 6 = extreme blame*).

Cause. Participants rated how much the agent was the cause of the outcome (*1 = not the cause at all, 6 = definitely the cause*).

Capacity. Participants rated how likely it was that the agent could have prevented the outcome (*1 = not likely at all, 6= very likely*).

Immoral character. Participants rated how good-bad, moral-immoral, trustworthy-untrustworthy the agent was (*1 = extremely good, moral, trustworthy; 6 = extremely bad, immoral, untrustworthy*), αs > .92.

Results and Discussion

We conducted a 2 (Omission) x 2 (Harm) x 3 (Scenarios) repeated measures Analyses of Variance (ANOVAs) on the main dependent variable, i.e., blame. The 3-way interaction turned non-significant, $F(2, 204) = 2.22$, $p = .11$, just as the Harm x Scenarios interaction, $F(2, 204) = 0.44$, $p = .65$, and the Omission x Harm interaction, $F(1, 102) = 0.12$, $p = .73$. However, there was a significant Omission x Scenario interaction, $F(2, 204) = 5.54$, $p = .005$. Contrast analyses for each of the three group of scenarios showed that there was a main effect of inaction for all set of scenarios, $F(1, 102) = 39.64$, $p < .001$(for scenarios A and D), $F(1, 102) = 5.09$, $p = .03$ (for scenarios B and E), $F(1, 102) = 28.97$, $p < .001$ (for scenarios C and F), qualifying greater blame for inaction (vs. action). Hence, all scenarios showed a similar pattern of results for Omission regardless of the Harm condition. The main effects of Harm and Scenarios are beyond the scope of this study. Therefore, we will report only the main effects of Omission.

An agent who did not take action was considered to be more blameworthy ($F(1, 104) = 34.88$, $p < .001$, $\eta_p^2 = 0.26$), more of a cause of the negative outcome ($F(1, 104) = 14.32$, $p < .001$, $\eta_p^2 = 0.12$), more capable of preventing the outcome ($F(1, 109) = 40.02$, $p < .001$, $\eta_p^2 =$

0.27), and more immoral ($F(1, 106) = 130.39$, $p < .001$, $\eta_p^2 = 0.56$) than one who did take action. Table 3 shows the means and standard deviations for each analysis.

Table 3. Means and standard deviations for each dependent variable of Study 1.1.

Means and Standard Deviations for each Dependent Variable		
	M (SD)	
DVs	Action	Inaction
Blame	2.07 (1.23)	2.46 (1.49)
Causality	1.96 (1.21)	2.75 (1.47)
Capacity	3.16 (1.27)	4.62 (1.27)
Immoral Character	2.55 (1.24)	4.50 (0.94)

Mediation

To investigate the potential psychological processes underlying the effect of omission on blame, we performed mediation analyses with bootstrapping (10,000 samples) using PROCESS on SPSS, model 4 (Hayes, 2013). The action condition was coded as 0, and the inaction condition as 1. Cause, capacity and immoral character were treated as mediators. Inaction increased blame indirectly through its effect on perceived causality (0.54, 95% CI = 0.27 to 0.84), capacity (0.17, 95% CI = 0.01 to 0.38), and immoral character (0.29, 95% CI = 0.05 to 0.55) (see Figure 2).

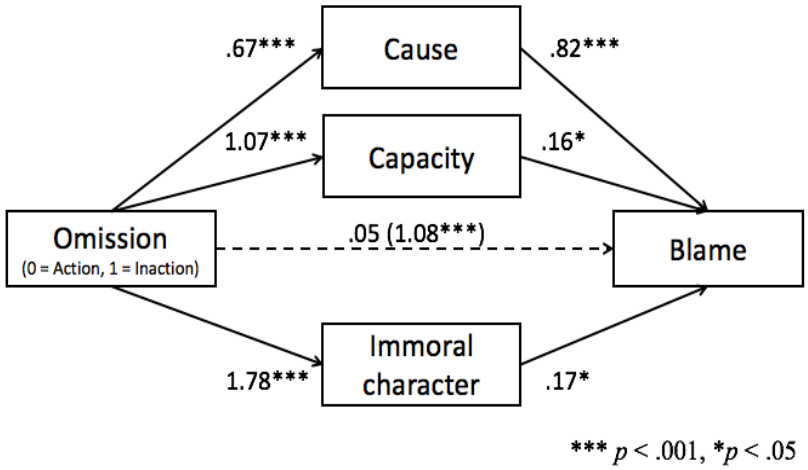

Figure 2. A parallel mediation analysis revealed that omission (dummy coded as 0 = action, 1 = inaction) affected blame judgments indirectly through its effect on judgments of causality, capacity, and immoral character.

Participants assigned greater blame to an agent who omitted help than to one who attempted but failed to help. This effect took place partly because they perceived the agent to have a greater causal role in the outcome and to be more capable of preventing it. Note, however, that the ratings of cause and capacity were overestimated. In both inaction and action conditions, the agent did not directly cause the outcome. Moreover, whereas taking action confirms the agent's incapacity of helping, omitting behavior logically should not make one more capable. This illusory judgment of capacity appears to be a byproduct of counterfactual thinking. Finally, the agents who omitted help were judged as more blameworthy in part because they were perceived as worse characters. This provides initial evidence of a person-based motivation to blame omission.

One limitation of this study was the lack of control for intentions in the scenarios, making it unclear whether the effect found for moral character could be explained by perceived valence of intentions. Moreover, throughout the scenarios, the agents showed different reasons for failing to take action (e.g., he thought there was nothing he could do; he thought it had nothing to do with him), which could have altered the blame judgments. We address these issues in Study 2.

Study 1.2

In Study 1.2, to clearly establish whether immoral character explains the effect of omission on blame, we held causality, intentions, capacity, obligation, and justification constant. We used two scenarios of Study 1.1 describing agents who failed to help by inaction (vs. action) yet did not actively cause the negative outcome. This time, we added situational constraints to strengthen control over potential extraneous variables. That is, in both conditions, we informed that the agents (1) were physically incapable of helping the victim; (2) intended to help; and (3) acknowledged their own incapacity of helping, hence providing a plausible justification for the agents to omit help. No specific information on obligation was provided. Based on the literature, we reasoned that blame would hardly be assigned in such extreme scenarios, unless other factors such as inferences of moral character would account for blame. This study tests this hypothesis. Finally, to control for the potential order effect of the dependent variables, participants were randomly assigned to assess either blame first or causality first. Hence, this study employed a 2 (Omission: action, inaction) x 2 (Order: blame first, causality first) x 2 (Scenarios) fully between-subjects design.

Method

Participants

We recruited 434 Japanese participants through an outsourcing service named CrowdWorks (241 females, M_{age} = 38.38, SD = 9.82). The sample size was determined by previous GPower analyses, which suggested 387 participants (η_p^2 = 0.02, 80% power, df = 1, groups = 8). There was no exclusion of participants.

Materials and procedures

We revised two scenarios from Study 1 which involved direct harm to the victim (for full scenarios, see Appendix 1.2). The first scenario described Nakamura, a man who had been practicing swimming across a lake and who, after a month of practice, could only get three-quarters across. On his 31st training day, Nakamura saw a man drowning on the far side of the lake and wanted to save him. With no time to call for help and with full knowledge that he could not swim across the lake, Nakamura decided either to swim to the man yet fail to reach him in time, or not to. In both conditions, the man drowned.

The second scenario described a runner who failed to rescue a child about to fall from a tree; the negative outcome was that the injury from the fall left the child with a lifelong sequelae. After reading either one of the scenarios, participants answered on 11-point scales:

Blame. Participants rated how much they blamed the agent for the man's death/child's injury (1 = *no blame at all, 11 = extreme blame*).

Cause. Participants rated how much Nakamura was the cause of the man's death/child's injury (*1 = not the cause at all, 11 = definitely the cause*).

Intentions to help. Participants rated the extent to which they thought Nakamura intended/desired to help the man/child (*1 = did not intend/desire at all, 11 = strongly intended/desired*), r (434) = .83, $p < .001$.

Justification. Participants rated the extent to which Nakamura's failure to help was inevitable/justifiable (*1 = not at all, 11 = definitely inevitable/justifiable*), r (434) = .79, $p < .001$.

Capacity. Participants rated how likely it was that Nakamura could have prevented the man's death/child's injury if he were to go through the same situation again (1 = *not likely at all, 11 = very likely*).

Obligation. Participants rated the extent to which Nakamura had an obligation to help the man/child (*1 = no obligation at all, 11 = extreme obligation*).

Immoral character. Participants rated how good-bad, moral-immoral, trustworthy-untrustworthy the agent was (*1 = extremely good/ moral/ trustworthy; 11 = extremely bad/ immoral/ untrustworthy*), $\alpha = .94$.

Results and Discussion

We conducted a 2 (Omission) x 2 (Order) x 2 (Scenarios) repeated measures Analyses of Variance (ANOVAs) on the main dependent variable, i.e., blame, with focus on the effect of Order. The main effect of Order was non-significant, $F(1, 426) = 1.04$, $p = .31$, just as the 2-way interactions, Omission x Order interaction, $F(1, 426) = 3.02$, $p = .08$, and Scenarios x Order interaction, $F(1, 426) = 0.20$, $p = .65$, and the 3-way interaction, $F(1, 426) = 0.11$, $p = .74$. Hence, the results suggested there was no significant difference in the ratings of blame depending on the order in which participants rated blame and causality.

As we aimed to collapse the analyses across scenarios, we proceeded by analyzing the effect of Omission and Scenarios. The main effect of Omission was significant, $F(1, 426) = 36.63$ $p < .001$, $\eta_p^2 = .08$, the main effect Scenarios was non-significant, $F(1, 426) = 0.36$, $p = .55$, and there was a significant 2-way Omission x Scenarios interaction, $F(1, 426) = 6.43$, $p = .01$, $\eta_p^2 = .02$. However, contrast analyses showed a similar effect of Omission for each scenario, with greater blame for inaction (vs. action) for the Swimmer scenario, $F(1, 426) = 6.03$, $p = .01$, $\eta_p^2 = .02$, and the Runner scenario, $F(1, 426) = 37.84$, $p < .001$, $\eta_p^2 = .08$. We conducted 3-ways ANOVAs for all other dependent variables and found a similar pattern of results. Therefore, we considered appropriate to disregard the effects of Scenarios and Order and report only the main effect of Omission, which was the focus of this study.

Participants judged an agent who did not take action (vs. who took action) to be more blameworthy for the outcome ($F(1, 426) = 36.63$ $p < .001$, $\eta_p^2 = 0.08$), to be more of a cause of the negative outcome ($F(1, 426) = 10.15$, $p = .002$, $\eta_p^2 = 0.02$), and to be less intentional of helping ($F(1, 426) = 150.70$, $p < .001$, $\eta_p^2 = 0.26$). Moreover, their failure to help was rated as less justifiable, ($F(1, 426) = 23.35$, $p < .001$, $\eta_p^2 = 0.05$), and they were perceived as more immoral ($F(1, 426) = 302.47$, $p < .001$, $\eta_p^2 = 0.42$). Finally, participants did not differentiate the agent's capacity of helping across conditions, ($F(1, 426) = 0.31$, $p < .58$), and they perceived greater obligation for the action condition ($F(1, 426) = 7.90$, $p = .005$, $\eta_p^2 = 0.02$). For means and standard deviations for each dependent variable, see Table 4.

Table 4. Means and standard deviations for each dependent variable of Study 1.2.

Means and Standard Deviations for each Dependent Variable		
	M (SD)	
DVs	Action	Inaction
Blame	1.68 (1.34)	2.73 (2.23)
Causality	1.85 (1.48)	2.42 (1.93)
Intentions to help	9.48 (1.98)	6.73 (2.60)
Justification	8.94 (2.30)	7.71 (2.69)
Immoral Character	2.90 (1.66)	5.80 (1.83)
Capacity	4.24 (2.27)	4.34 (2.36)
Obligation	5.36 (2.57)	4.67 (2.51)

Mediation

Mediation analyses revealed that Inaction (dummy coded as action = 0 and inaction = 1) affected blame indirectly through its effect on perceived causality (0.24, 95% CI = 0.10 to 0.42), justification (-0.23, CI 95% CI = -0.41 to -0.10), and immoral character (0.67, 95% CI = 0.29 to 1.06), whereas indirect effects were non-significant for intentions (-0.25, 95% CI = -0.51 to 0.00), capacity (0.00, 95% CI = -0.01 to 0.02) and obligation (-.04, 95% CI = -0.09 to 0.00) (see Figure 3).

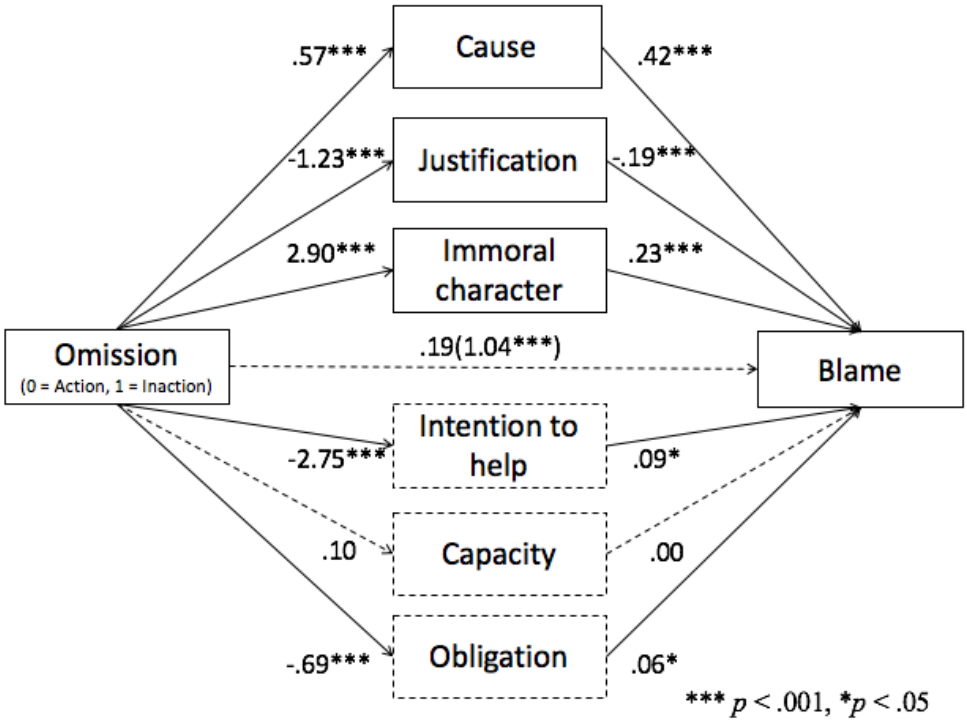

Figure 3. A parallel mediation analysis revealed that omission (dummy coded as 0 = action, 1 = inaction) affected blame judgments indirectly through its effect on judgments of causality, justification and immoral character. The indirect effects by intentions, capacity, and obligation were non-significant (represented by dotted lines).

We depicted an extreme scenario in which the agent who failed to help did not cause the outcome, was shown to explicitly desire to help and was incapable of helping, hence providing justification for inaction. The results replicated the omission effect on blame found in Study 1 with a more general sample. Moreover, although participants recognized that the agents in both conditions were similarly incapable of helping (i.e., non-significant main effect of Omission on capacity), they judged failure to help by omission (vs. failed attempt) as more causal and as less justifiable. This puzzling result suggests unreasonable harshness in judging omission. As shown in the mediational analyses, these judgments explained the effect of omission on blame. Finally, once again, immoral character was also a significant mediator. The inclusion of intentions to the model ruled out the hypothesis that the mediational role of moral character found in Study 1 could be explained by intentions.

Study 1.3

We aimed to replicate the moral character effect on blame through an experimental design. Moreover, based on the person-based approach to moral judgments, bad agents should receive greater blame for omitting help than good agents. Hence, we sought to examine this potential interactive effect of omission and moral character disposition.

Study 1.2 depicted agents who were incapable of helping. This time, we portrayed the agent as having a 50% chance of succeeding in helping the victim, that is, either failure or success in helping would take place by chance rate. With this modification, we aimed to identify whether capacity would be overestimated for the omission condition. This study consisted of a 2 (Omission: inaction, action) x 3 (Moral Character: immoral, moral, neutral) x 2 (Scenarios) between-subjects design.

Method

Participants

We recruited 610 Japanese participants through CrowdWorks for the price of 100 yen. We explicated that participants would only be compensated if they had not participated in any of our previous studies. They were also required to complete the entire survey. We excluded 54 participants who had participated in our previous studies, and 48 participants who failed to complete the study. For the analyses, we also excluded 2 participants who failed the attention check items, 9 participants who took more than 30 minutes to complete the study, and 1 participant who took less than 2 minutes to complete the study. After the exclusion procedure, we had 496 participants in this study (276 females, Mage = 44.93, SD = 12.59). They took an average of 8.96 minutes to complete the survey (maximum of 28.55 minutes, and minimum of

2.67 minutes). Previous sample size determination analyses required a minimum of 476 participants (η_p^2 = 0.02, 80% power, *df* = 2, groups = 12).

Materials and procedures

Participants were randomly assigned to one of the three conditions varying in moral disposition of the agent. In the "immoral" condition, the agent was indifferent to other people's feelings, gossiped, feigned sickness to skip work, and had been cheating on his wife for ten years. In contrast, in the "moral" condition, the agent was an empathic man, sensitive to other people's feelings, discreet, never missed work, and had been faithful to his wife for ten years. Finally, in the "neutral" condition, the agent was an ordinary man who spent his time doing research (see Appendix 1.3 for full description). In a pilot study (n = 122), participants rated the agents as immoral (*M* = 1.55, *SD* = 0.82), moral (*M* = 10.48, *SD* = 0.72), and neutral (*M* = 7.44, *SD* = 2.36), *F* (2, 119) = 371.12, $p < .001$, η_p^2 = 0.86, consistent with the respective condition.

Participants were also randomly assigned to read either the Swimmer or the Child scenario (see Appendix 1.3). The Swimmer scenario described Nakamura, a man who had been practicing swimming across a lake without taking a break. Nakamura did not practice regularly, so his performance was inconsistent. In 20 practice sessions, Nakamura had been able to swim across the lake only 10 times. One day Nakamura was swimming, when he noticed a man drowning on the far side of the lake. There was no time to get help. Here again, Nakamura either failed to help by omission or by failed attempt. After reading the scenarios, participants rated on 11-point scales:

Blame. Participants rated how much they blamed Nakamura for the man's death/ the child's injury (*1* = no blame at all, *11* = extreme blame).

Cause. Participants rated how much they thought that Nakamura was the cause of the man's death/ the child's injury (*1* = not the cause at all, *11* = definitely the cause).

Intentions. Participants rated how much they thought that Nakamura intended to help the man/ the child (*1* = did not intend at all, *11* = strongly intended to help) and that Nakamura desired to help the man/the child (*1* = did not desire at all, *11* = strongly desired to help), r (496) = .88, $p < .001$.

Justification. Participants rated how much Nakamura's failure to help was inevitable (*1* = not inevitable at all, *11* = definitely inevitable), and how much Nakamura's failure to help was justifiable (*1* = not justifiable at all, *11* = definitively justifiable), r (496) = .74, $p < .001$.

Capacity. Participants rated how likely Nakamura would have been able to prevent the man's death/ the child's injury (*1* = not likely at all, *11* = very likely); how much they thought that Nakamura would have been able to save the man / the child if only he had tried harder? (*1* = don't think like that at all, *11* = strongly think like that); and how much they thought the man's death/ the child's injury could have been prevented if only Nakamura had acted differently (*1* = don't think like that at all, *11* = strongly think like that), (α = .83).

Obligation. Participants rated how much they thought that Nakamura had an obligation to help the man/ the child (*1* = no obligation at all, *11* = extreme obligation).

Moral character. Participants rated how good-bad, moral-immoral, trustworthy-untrustworthy the agent was (*1 = extremely good/ moral/ trustworthy; 11 = extremely bad/ immoral/ untrustworthy*), α = .96.

Social distance. To understand the potential social consequences of the failure to help, we included social distance items (Bogardus, 1933). Participants rated the desirability of having the

agent as a friend, family member, or co-worker (*1 = not desirable at all, 11 = extremely desirable,* $\alpha = .92$).

Results and Discussion

We conducted a 3 (Moral disposition) x 2 (Omission) x 2 (Scenarios) Analyses of Variance (ANOVAs) on the blame scores. The results of 3-way interaction turned out to be non-significant, $F(2, 484) = 0.17$, $p = .85$, just as the 2-way Moral disposition x Scenarios interaction, $F(2, 484) = 0.42$, $p = .66$. There was a significant 2-way Omission x Scenario interaction, $F(1, 484) = 10.27$, $p < .001$, $\eta_p^2 = 0.02$. This interaction revealed that inaction (vs. action) led to greater blame for both scenarios, however even more so for the Runner scenario, that is, the overall pattern of the scenarios was the same. The main effect of Scenario was also non-significant, $F(1, 484) = 3.10$, $p = .08$, $\eta_p^2 = 0.01$. Consistently, the same analyses for the ratings of the remaining variables (except the "immoral character" scores) showed that there were no meaningful significant interactions involving the Scenario factor. Therefore, with the exception of the immoral character scores, we will only report the main effects of Omission and Moral Disposition (see Table 5 for means and standard deviations).

Table 5. Means and standard deviations for each dependent variable of Study 1.3

Means and Standard Deviations for each Dependent Variable			
		M (*SD*)	
DVs		Action	Inaction
Immoral Character	Moral	2.72 (2.05)	4.46 (2.48)
	Neutral	3.79 (2.24)	6.40 (2.04)
	Immoral	8.48 (2.30)	9.19 (1.84)
	Mean	4.82 (3.28)	6.67 (2.83)
Blame	Moral	1.85 (1.53)	3.62 (2.43)
	Neutral	2.02 (1.65)	4.15 (2.76)
	Immoral	2.91 (2.47)	5.38 (2.89)
	Mean	2.23 (1.94)	4.37 (2.79)
Causality	Moral	1.94 (1.77)	3.14 (2.18)
	Neutral	2.16 (1.72)	3.69 (2.40)
	Immoral	2.57 (2.14)	4.53 (2.57)
	Mean	2.21 (1.88)	3.79 (2.44)
Intention to help	Moral	8.88 (1.49)	6.43 (2.22)
	Neutral	8.51 (1.31)	5.64 (2.01)
	Immoral	7.15 (1.72)	4.41 (2.04)
	Mean	8.24 (1.67)	5.50 (2.23)
Justification	Moral	9.48 (1.77)	8.20 (2.22)
	Neutral	9.02 (1.82)	7.46 (2.15)
	Immoral	8.70 (1.95)	6.93 (2.40)
	Mean	9.09 (1.86)	7.52 (2.30)
Capacity	Moral	4.40 (2.20)	5.66 (2.31)
	Neutral	4.73 (1.99)	6.01 (2.08)
	Immoral	5.24 (1.99)	6.48 (2.11)
	Mean	4.76 (2.09)	6.05 (2.18)
Obligation	Moral	4.57 (2.72)	4.21 (2.56)
	Neutral	4.64 (2.55)	5.36 (2.71)
	Immoral	5.61 (2.83)	5.45 (2.66)
	Mean	4.91 (2.73)	5.04 (2.70)

Manipulation check

Participants assigned greater moral character to the agent with moral (vs. immoral and vs. neutral) disposition (F (2, 484) = 242.96, $p < .001$, $\eta_p^2 =$ 0.50). Moreover, the Omission x Moral Disposition interaction yielded significant results. Pairwise comparisons suggested that inaction

(vs. action) was informed greater immoral character for all levels of Moral disposition (all *p*s < .05), yet less so for the immoral condition.

Blame

One purpose of this study was to reveal whether the effect of omission on blame depends on the moral disposition of the agent. There was no evidence of such interaction (F (2, 484) = 0.94, $p < .39$, $\eta_p^2 = 0.00$). Agents received greater blame when they chose inaction (vs. action), F (1, 484) = 101.58, $p < .001$, $\eta_p^2 = 0.17$, and the more immoral they were, F (2, 484) = 15.00, $p < .001$, $\eta_p^2 = 0.06$ (see Figure 4).

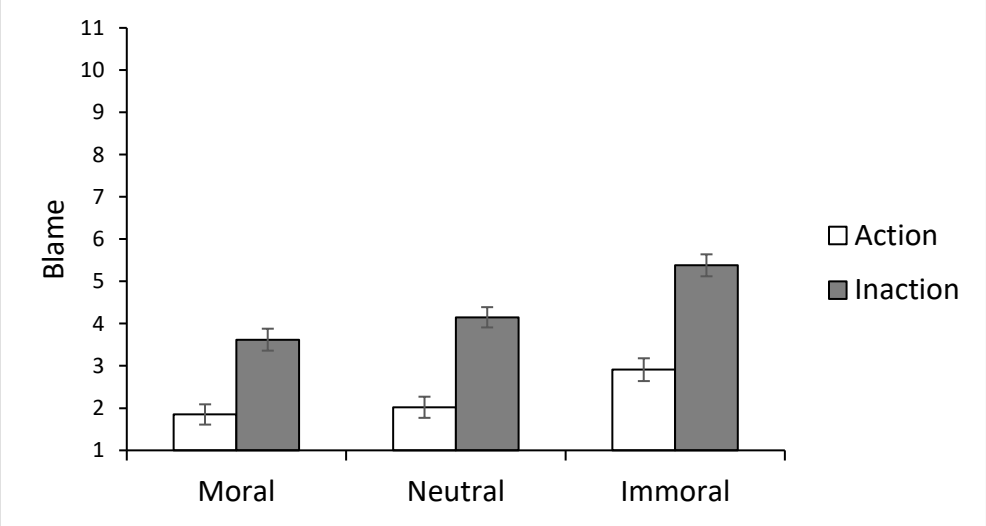

Figure 4. Ratings of blame for failure to help for action vs. inaction conditions across each moral disposition condition. There was a significant main effect of the omission, with greater blame for the inaction (vs. action) condition. The main effect of moral character disposition also turned significant, with greater blame for individuals of immoral (vs. moral and neutral) disposition. The two-way interaction was non-significant. Bars represent standard errors.

Mediators

When the agent chose inaction (vs. action), participants gave higher ratings for causality (F (1, 484) = 64.32, $p < .001$, $\eta_p^2 = 0.12$), lower ratings of intentions (F (1, 484) = 270.58, $p < .001$, $\eta_p^2 = 0.36$) and justification (F (1, 484) = 69.10, $p < .001$, $\eta_p^2 = 0.13$), and higher ratings

for immoral character (F (1, 484) = 74.37, $p < .001$, $\eta_p^2 = 0.13$). Moreover, agents who chose inaction (vs. action) were perceived to be more capable of helping if the situation were to happen a second time, F (1, 484) = 43.32, $p < .001$, $\eta_p^2 = 0.08$, despite the factual information that the agent had a 50% chance of succeeding in helping. Obligation did not differ across inaction and action conditions (F (1, 484) = 0.03, $p = 0.86$, $\eta_p^2 = 0.00$).

More immoral disposition also led to greater causality (F (2, 484) = 8.54, $p < .001$, $\eta_p^2 = 0.03$), lower intentions (F (2, 484) = 43.85, $p < .001$, $\eta_p^2 = 0.15$) and justification (F (2, 484) = 9.53, $p < .001$, $\eta_p^2 = 0.04$), and greater capacity (F (2, 484) = 5.75, $p = .003$, $\eta_p^2 = 0.02$) and obligation (F (1, 484) = 7.01, $p = 0.01$, $\eta_p^2 = 0.03$).The effect of moral disposition on these variables suggests a pervasive motivation to condemn immoral characters. It also suggests an intricate process of interinfluence among these moral judgments. However, to understand the single contribution of each of these components on blame, we judged worthwhile to conduct a parallel mediation analysis.

Mediation

Inaction (dummy coded as 0 = action, 1 = inaction) once again affected blame judgments indirectly through its effect on perceived causality (0.85, 95% CI = 0.59 to .1.15), intentions (0.32, 95% CI = 0.10 to 0.54), justification (0.43, 95% CI = 0.24 to 0.67), and immoral character (0.09, 95% CI = 0.01 to 0.18). The mediation by capacity (0.09, 95% CI = -0.02 to 0.22) and obligation (0.01, 95% CI = -0.02 to 0.06) were non-significant (see Figure 5).

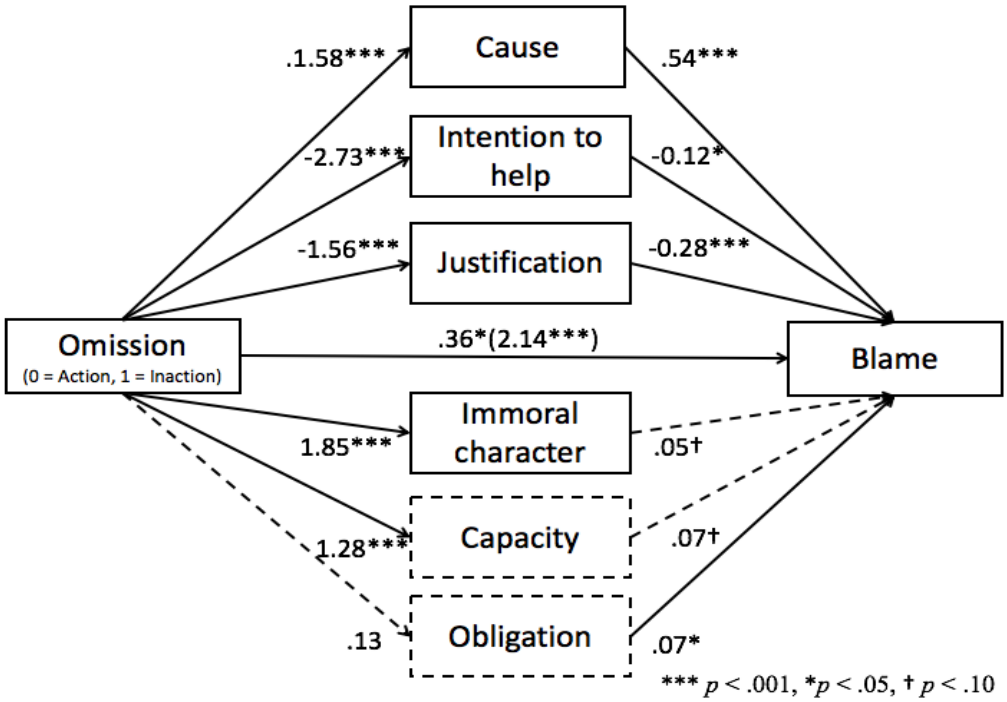

Figure 5. A parallel mediation analysis revealed that omission (dummy coded as 0 = action, 1 = inaction) affected blame judgments indirectly through its effect on judgments of causality, intentions, justification, and immoral character. The indirect effects by capacity and obligation were non-significant (represented by dotted lines).

Social distance

Finally, having the agent as a friend, family member, and coworker was less desirable in the inaction (vs. action) condition ($F(1, 484) = 85.23$, $p < .001$, $\eta_p^2 = 0.15$) and the more immoral their disposition ($F(2, 484) = 290.71$, $p < .001$, $\eta_p^2 = 0.55$). The two-way interaction was significant, $F(2, 484) = 13.81$, $p < .001$, $\eta_p^2 = 0.05$ (see Figure 6). Pairwise comparisons revealed that, for neutral and moral agents, there were social consequences of omission, that is, not taking action (vs. action) mitigated desirability (*p*s < .001). However, when the agent was immoral, the omission manipulation did not affect desirability ($p = .12$).

Figure 6. Ratings of desirability of interacting with the agent for action vs. inaction conditions across each moral disposition condition. We observed a significant two-way interaction. The main effect of the omission manipulation was significant for the moral and neutral moral disposition conditions but was non-significant for the immoral condition. Bars represent standard errors.

This study makes three points. First, there was an omission effect on blame regardless of the agent's moral disposition. Second, the results replicated the mediational role of causality, justification, and immoral character. Unlike in Study 1.2, intention was a significant mediator. This likely occurred because, in this study, we did not provide explicit information on the agent's intentions, allowing participants to infer intentions based on the action information. We also demonstrated an overestimation of capacity for omissions, suggesting that participants engaged in a biased reasoning of "capable until proven wrong." However, this judgment did not explain judgments on blame. Finally, we found that choosing omission mitigated the desire to interact with the agents, at least when they displayed a neutral or moral disposition. Moreover, when the agents were immoral, taking action (vs. inaction) did not make interaction with them more desirable. This interesting result reveals the importance of perceived morality in determining social interaction. It also suggests that immoral individuals may not be socially redeemed by acting morally.

General Discussion for Study 1

In three studies, agents who chose to omit help (vs. who attempted to help) were deemed more blameworthy for the negative outcome that followed their failure to help. This omission effect took place in situations that involved both direct and indirect harm (Study 1.1), remained even when it was impossible for the agent to succeed in helping, hence logically justifying their omission (Study 1.2), and was independent of previous information on the moral disposition of the agent (Study 3). The choice to omit help also affected further social decisions, such as the desire to interact with the agent (Study 1.3).

In typical omission bias studies, a confound for the preference of omission over commissions is the status quo bias. For instance, people prefer to not vaccinate a child when the vaccine can cause death versus vaccinate, even though vaccination reduces the risk of harm overall (Ritov & Baron, 1990). An argument for this preference is that people show a status quo bias, that is, a preference for maintaining the current state of affairs (Samuelson & Zeckhauser, 1988). Our findings rule out the status quo bias. In the present study, people showed preference (i.e., reduced blame) for individuals who attempted to help in comparison with individuals who omitted help, although omitting help represented the status quo state.

As discussed in the Chapter 1, the action-inaction effect on blame is typically explained by perceptions of causality. Our results are congruent with this explanation. It is puzzling, however, that the increased perception of causality for omission was not grounded on rational evidence, that is, there is no logical explanation to suppose that an individual who omitted to help is more causal of an outcome than an individual who attempted to help and failed.

Moreover, we also found an interesting augmentation of judgments of intention to help and justifiability, despite providing transparent information regarding these elements in the

scenarios. Previous research showed that when there is transparency of the decision-making to omit, the action-inaction effect on moral judgments disappears (DeScioli et al.'s (2011). In Study 1.2, the transparency was not sufficient to dissipate the action-inaction effect on blame. Despite the transparency, our results showed that an individual who chose to omit help (vs. who chose to act and failed) was perceived as a greater cause of the outcome, as having less intentions of helping and having less justification for the failure to help. This result is not obvious, because, based on the nature of the scenario, participants could have reasoned that there was nothing the agent could do to change the outcome, and hence alleviated these judgments. We considered that motivated reasoning may have altered these moral judgments (Alicke, 2000). In this light, the unfavorable evaluation of the agent's choice of omission might have led participants to exaggerate the agent's control over the negative outcome. This motivation to blame omission seems sensible when considering that helping others is typically reinforced across societies (Graham et al., 2013; Shweder et al., 1997), and that humans have evolved through cooperation with kin and reciprocation with people unrelated to them (Barret et al., 2002; Boster et al., 2001). Such ingrained moral heuristics may have led participants to inflate perceptions of causality, intentions to help and justifiability.

Naturally, the extent to which people are blamed for omitting help should depend on a variety of factors. The current studies are limited to situations of blame for choosing to omit help. Future studies may explore several potential factors that may influence this process of blame, such as the social roles, varying mental states, whether the situation is an emergency, the number of bystanders, individual differences, and the cost of helping (Dovidio et al., 2006; Haidt & Baron, 1996; Latané & Darley, 1968; Moussaïd & Trauernicht, 2016). It is likely, for instance, that blame for failure to help should be attenuated or inexistent in cases in which there is cost for helping (Bode et al, 2015). Another interesting venue for investigation is the potential actor-

observer effect (Jones & Nisbett, 1972) in judgments of blame for omission, with the possibility of individuals blaming omission to a lesser extent when they are the actors, rather than observers, due to potential asymmetries in reason perceptions, for instance (see Malle et al., 2007).

Finally, our key finding was that immoral character was consistently a significant mediator of the omission effect on blame, even after taking causality, intentions, justification, capacity, and obligation into account. This was an important methodological aspect of this work. The results supported the "people as intuitive lawyers" metaphor by showing that individuals blame based on important cognitive factors such as causal ascriptions, intentions and justification. It also supported the "person as intuitive virtue theorists" metaphor, showing that individuals were motivated to blame the immoral character revealed by the omission choice.

Curiously, omission received greater blame even when it represented a safer behavior. In Study 1.2, not taking action when it is known to be futile would conserve resources (e.g., energy spent trying to save the drowning man) and potential risk to the agent. Previous research reports a dissociation between moral judgments of acts and persons (Tannenbaum et al., 2011; Uhlmann et al., 2014). For instance, in Uhlmann et al.'s (2013) study, a hospital administrator who spent $2 million on hospital equipment and saved 500 lives (instead of funding a single life-saving operation for a little boy) was perceived as more deficient in moral character, despite making the more pragmatic and praiseworthy decision. Our results also seem indicative of this dissociation, as participants evaluated the agent who took the pragmatic decision to omit help as more immoral and more deserving of blame.

This study contributes to our understanding of the psychological mechanisms underlying blame for omitting help. Choosing to omit help is considered more blameworthy than attempting to help and failing, because an agent who omits help is perceived to be more causal, less intentional of helping, having less of a justification to help, and more of an immoral person. Our

findings add to the literature of the person-based approach to moral judgments by extending the character explanation to the context of blame for failure to help. There are also real-life implications to this study. Our results suggest that individuals blame omission of help through motivated processes, which may lead to unreasonable and unfair judgments.

Summary. Study 1 was composed of a set of three studies. We demonstrate that agents who chose to omit help (vs. who took action to help yet failed) were deemed more blameworthy for the negative outcome that followed their failure to help. This omission effect was found in situations that involved both direct and indirect harm (Study 1.1). It remained even when it was impossible for the agent to succeed in helping, hence logically justifying their omission (Study 1.2), and was independent of previous information on the moral disposition of the agent (Study 1.3). Importantly, the effect was mediated by inferences of immoral character, suggesting a motivation to blame immoral individuals. Moreover, perceptions of causality, intentions, and justification also consistently explained the omission effect on blame. This study provides empirical evidence that people blame individuals who omit help in part because they perceive them to be particularly immoral. This finding was replicated across three studies even when taking into account other elements that explain blame (Malle et al., 2014).

Chapter 3 - Study 2. Mental states and assignment of moral character

Blame for negative outcomes is determined by the agent's behavior (action or inaction) and the mental states of the agent. Study 1 showed evidence that people blame an agent who chooses to omit help in part because of inferences of immoral character. In Study 2 and Study 3, I examine whether the effect of mental states on blame is determined in part by inferences of the agent's moral character. In Study 2 of this thesis, I examined the first proposed path – that is, whether bare intentions are indicative of moral character.

Intention is a mental representation of a planned act predicted to produce a desired and believed outcome (Cushman, 2008). Especially when making moral judgments, we seem to care a great deal about what goes on inside other people's minds (Gray et al., 2012). For instance, ample evidence shows that greater intentionality of a committed immoral act is associated with harsher moral judgments, such as increased perception of harmfulness, blameworthiness, wrongness, punishment, and causality (Alicke, 1992; Ames & Fiske, 2013; Cushman, 2008; Monroe & Malle, 2017; Nadler, 2014; Young & Saxe, 2011). Furthermore, such effect of intentions takes place even in the absence of tangible consequences. Specifically, Cushman (2008) found that people judged acts that were intended as more blameworthy than unintended ones, even when none of the acts produced any outcome. Overall, these studies indicate that intentions have a fundamental role in the process of judging moral *acts* and are taken into consideration independently of associated outcomes.

The same seems to be true for the process of judging moral *persons*, which is the main interest of this study. Evidence in literature suggests that undesired thoughts or intentions by themselves lead to inferences of immoral character. For example, participants inferred greater immoral character of an agent who performed racial slur versus physical assault, even though the

latter was judged as a greater immoral act (Uhlmann et al., 2013). Likewise, hitting a cat was more informative of one's immoral character than hitting one's girlfriend (Tannenbaum et al., 2011). Possibly, the assumed morbid mental states underlying using racial slur and maltreating animals were more informative of immoral character than the factual harmfulness of the acts. Further works show a more straightforward association of mental states and moral character evaluations. For instance, studies by Cohen and Rozin (2001) found that inner states of disliking one's own parents or fantasizing about an affair (unfollowed by actual behavior) were informative enough to make a person a worse character, at least among Protestants. Other studies revealed that agents were judged as immoral even when they were not the cause of harm, to the extent that they experienced pleasure at others' harm (Gromet et al., 2016), and desired the harm to happen for an unrelated reason, i.e., "wicked desires" (Inbar et al, 2012). Similar to Cushman's (2008) account but in the context of person-based judgments, these studies indicate that the information of what is in an agent's mind is by itself sufficient for people to judge this agent to be moral or immoral.

Altogether, the literature showed that intentions matter for judgments of moral *acts*. It also indicates that intentions may affect judgments of moral *character*, as such effect was evident for other mental states (e.g., motives, attitudes). Hence, in the present study, we aimed to directly test the effect of intentions on moral character evaluations. Moreover, we also distinguished between intentions which are followed versus unfollowed by action (i.e., unfulfilled versus fulfilled intentions). We reasoned that, by analyzing intentions that *were not acted out,* we could obtain a clearer effect of the intentions themselves on moral character evaluations. This is also an important feature of our study, as most of past research has explored intentions in the context of acts that are committed intentionally or unintentionally (Ames & Fiske, 2013; Cushman, 2008; Young & Saxe, 2011).

In addition to the investigation of unfulfilled intentions, the second goal of this study was to explore a possible positive-negative asymmetry in its effect on moral character evaluations. The negativity bias is a tendency in many psychological processes, such as attention, memory, learning, and impression formation (Kensinger et al., 2006; Klein, 1991; Öhman et al., 2001; Peeters & Czapinsky, 1990; Pizarro et al., 2003; Reeder & Brewer, 1979). The literature on moral character judgments also reveals a negativity bias. For instance, doing a wrong had a greater impact in person evaluation despite the compensation of doing two rights (Riskey & Birnbaum, 1974), dishonest behavior predicted character evaluation to a greater extent than honest behavior (Skowronski & Carlson, 1987), and immoral (versus moral) attribution of character was less influenced by situational demands (Reeder & Spores, 1983). Displaying immoral behavior was more telling of one's moral character than moral behavior. Likewise, we expected that negative unfulfilled intentions would show a greater impact on moral character evaluations compared to positive ones.

Finally, we searched for potential explanations to the predicted effect. Our first mediator was emotions, based on the vast literature indicating that an evaluator's affective state strongly influences subsequent moral judgments (Avramova & Inbar, 2013; Forgas & Bower, 1987; Haidt, 2001; Ugazio et al., 2012). As we considered the asymmetry in our scope, we searched for both negative and positive emotions. With respect to negative emotions, previous research shows that experiencing anger and disgust, e.g., moral outrage (Salerno, & Peter-Hagene, 2013) motivates judgments of punishment and unfairness (Carlsmith et al., 2002; Goldberg et al., 1999; Mullen, & Skitka, 2006). With respect to positive emotions, "other-praising" affective states like liking and admiration are associated with witnessing moral behavior (Haidt, 2003), and they are also likely to affect further moral judgments. For instance, in a study by Bocian et al. (2018), participants who were induced to like a target rated this target with a more positive moral character.

Concerning admiration, we did not find a likewise straightforward evidence of its effect on judgments of moral character. However, studies show that feeling admiration towards performers of moral acts lead to increased reported inspiration (Van de Ven et al., 2018) and the wish to better oneself (Algoe & Haidt, 2009). It seems sensible that underlying such inspiration and wish to better oneself is the assessment of great moral character of the moral agent. Altogether, previous research suggests that emotions driven by moral situations, either negative (i.e., anger and disgust) or positive (i.e., liking and admiration), color following moral judgments. With our first mediator, we aimed to capture this affective role in our study.

Lastly, we also examined the potential role of a cognitive-based mediator, namely informativeness. Early attribution theories have long verified that certain acts are more informative of character than others. Acts performed consistently are more likely to diagnose character, as they seem to elucidate one's pattern of behavior. Similarly, acts that are not performed by most people (i.e., less consensual) are more likely to be driven by individual, rather than social motives (Jones & Davis, 1965; Kelley, 1967). Taken that consistency and perceived consensus of an *act* are important informational cues for character inference, we predicted that such frequency-based perceptions would also be relevant in the domain of *mental states*. In other words, *intentions* would predict moral character evaluations to the extent that they appeared to be consistent and distinct from others, that is, informative of one's dispositions. Again, we expected this to be true especially for negative intentions, drawing on the evidence that immoral (vs. moral) behaviors tend to be more diagnostic of character due to its low frequency (Mende-Siedlecki et al., 2013; see also Chakroff &Young, 2015). To the best of our knowledge, this is the first study to explore the role of an intention's informativenes on moral character evaluations.

In sum, in the present study, we aimed to investigate the effect of intentions unfollowed by action on judgments of moral character. In exploring this effect, we hypothesized a positive-

negative asymmetry characterized by stronger effects for negative (vs. positive) intentions. Furthermore, to uncover underlying psychological processes, we tested the potential mediation by affective processes (i.e., emotions) and cognitive processes (i.e., informativeness).

Study 2.1

Method

Participants

Japanese students from a psychology course (N = 91) participated in the survey (83 men, M_{age} = 19.91, *SD* = 1.10).

Materials and procedures

We prepared four scenarios of agents who intended to engage in either positive or negative actions. The agent with positive intentions intended to (1) donate money or (2) help a classmate with their study, whether the agent with negative intentions intended to (3) hit a colleague or (4) leave the train after noticing a child with Down syndrome. In a preliminary study (n = 68), participants rated how bad-good the actions were (*-5 = bad, +5 = good*). Results showed that the positive scenarios represented good acts (*M* = 3.30, *SD* = 1.82 for the donation scenario; *M* = 3.57, *SD* = 1.16 for the helping scenario), and negative scenarios represented bad acts (*M* = -2.92, *SD* = 2.10 for the hit scenario; *M* = -1.68, *SD* = 2.04 for the prejudice scenario). We manipulated whether the agent fulfilled or not their intentions to help/harm. Hence, this study consisted of a 2 (Intentions: positive, negative) x 2 (Fulfillment: unfulfilled, fulfilled) x 2 (Scenarios) mixed design, with the latter as the within-subjects variable.

After reading either two positive or two negative scenarios (counterbalanced), participants rated:

Moral character. Participants rated how immoral-moral and bad-good the agent was, *r*s (91) > .74, *ps* < .001 (-5 = *extremely immoral-bad, +5 = extremely moral-good*)

Emotions towards the agent. Participants rated how intensely they felt negative emotions (anger and disgust, *rs* (91) > .70, *ps* < .001) and positive emotions (respect and

attraction, *rs* (91) > .71, ps < .001) towards the agent (*1 = no feeling whatsoever, 7 = extreme feelings of designated emotion*). A general score of emotions was calculated by subtracting negative emotions scores from positive emotions scores. Higher ratings represented greater positive emotions.

Results and Discussion

Moral character

Firstly, we conducted a three-way Analysis of Variance (ANOVAs) for repeated measures on the moral character scores. There was a significant main effect of Fulfillment ($F(1, 87) = 5.53$, $p = .02$, $\eta_p^2 = 0.06$), a significant main effect of Valence ($F(1, 87) = 105.60$, $p < .001$, $\eta_p^2 = 0.55$), and a non-significant main effect of Scenarios ($F(1, 87) = 0.39$, $p = .54$). The Valence x Fulfillment interaction was significant ($F(1, 87) = 35.26$, $p < .001$, $\eta_p^2 = 0.29$), and so was the Scenario x Fulfillment interaction ($F(1, 87) = 7.64$, $p = .001$, $\eta_p^2 = 0.08$), whereas the Scenario x Valence interaction was non-significant ($F(1, 87) = 0.52$, $p = .64$). These main effects were qualified by a significant three-way interaction, $F(1, 87) = 14.61$, $p < .001$, $\eta_p^2 = 0.15$. Figure 7 shows the means for moral character evaluation across each condition of fulfillment for each scenario.

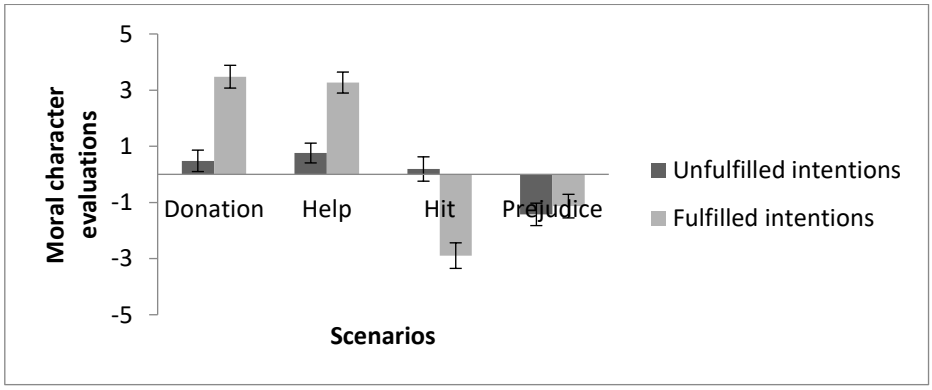

Figure 7. Moral character ratings across valence and fulfillment conditions. We observed a significant 3-way interaction. The difference between the unfulfilled and fulfilled conditions was non-significant only for the prejudice scenario. Bars represent standard errors.

For the positive scenarios, there was a similar pattern of the effect of fulfillment on moral character evaluations. An agent who fulfilled the intention to donate (M = 3.47, SD = 2.10) was perceived to be more moral compared to an agent who did not fulfill their intentions (M = 0.48, SD =1.87), $F(1, 87) = 29.04$, $p < .001$, $\eta_p^2 = 0.25$ (Bonferroni correction was used for all pairwise comparisons). Likewise, an agent who fulfilled the intention to help (M = 3.27, SD = 1.94) was perceived to be more moral compared to an agent who did not fulfill their intentions (M = 0.76, SD = 1.73), $F(1, 87) = 23.83$, $p < .001$, $\eta_p^2 = 0.22$.

For the negative scenarios, the effect of fulfillment showed different patterns between the scenarios. Similar to the positive scenarios, an agent who intended to hit a classmate was perceived to be more immoral when they fulfilled their intentions (M = -2.89, SD = 2.08) compared to when they did not fulfill it (M = 0.19, SD = 1.88), $F(1, 87) = 18.55$, $p < .001$, $\eta_p^2 = 0.17$. On the other hand, an agent who intended to act out on their prejudice was perceived to be immoral to a similar extent regardless of fulfilling (M = -1.13, SD = 1.92) or not (M = -1.42, SD = 1.74) their intentions, $F(1, 87) = 0.30$, $p = .58$.

Emotions

The same procedure followed for the scores of emotions. For this measure, the main effect of Fulfillment reached significance only at marginal level ($F(1, 86) = 3.16$, $p = .08$, $\eta_p^2 = 0.04$). There was a significant main effect of Valence ($F(1, 86) = 64.15$, $p < .001$, $\eta_p^2 = 0.43$), and a non-significant main effect of Scenarios ($F(1, 87) = 1.05$, $p = .30$). The Valence x Fulfillment interaction was significant ($F(1, 87) = 9.46$, $p = .003$, $\eta_p^2 = 0.09$). The Scenario x Fulfillment interaction was non-significant ($F(1, 87) = 0.01$, $p = .91$), whereas the Scenario x Valence interaction was significant ($F(1, 87) = 13.02$, $p = .001$, $\eta_p^2 = 0.13$). These main effects were

qualified by a significant three-way interaction, $F(1, 87) = 4.76$, $p = .03$, $\eta_p^2 = 0.05$. Figure 8 shows the means for emotions across each condition of fulfillment for each scenario.

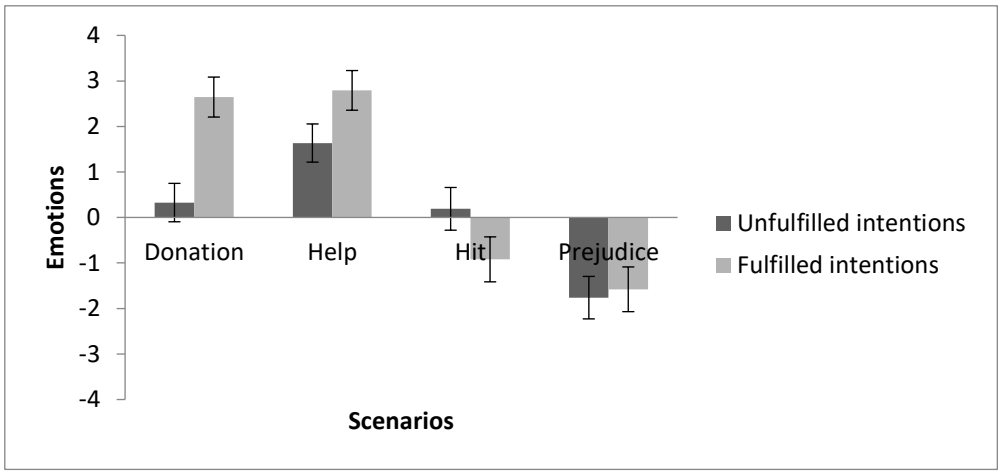

Figure 8. Emotions ratings across valence and fulfillment conditions. We observed a significant 3-way interaction. Bars represent standard errors.

For the positive scenarios, there was a similar pattern of the effect of fulfillment on moral character evaluations. Participants felt greater positive emotions towards an agent who fulfilled the intention to donate ($M = 2.64$, $SD = 2.28$) compared to one who did not fulfill their intentions ($M = 0.33$, $SD = 2.06$), $F(1, 86) = 14.50$, $p < .001$, $\eta_p^2 = 0.14$ (Bonferroni correction was used for all pairwise comparisons). At a marginal level, an agent who fulfilled the intention to help ($M = 2.79$, $SD = 2.26$) also received greater positive emotions compared to the agent who did not fulfill their intentions ($M = 1.63$, $SD = 2.05$), $F(1, 86) = 3.65$, $p = .06$, $\eta_p^2 = 0.04$.

For the negative scenarios, the effect of fulfillment was non-significant for both scenarios. Participants felt negative emotions towards an agent who intended to hit a classmate to a similar extent regardless of fulfilling ($M = -0.92$, $SD = 2.26$) or not ($M = 0.19$, $SD = 2.04$) their intentions, $F(1, 86) = 2.66$, $p = .11$. A similar pattern was found for an agent who intended to behave in a prejudiced way ($M = -1.57$, $SD = 2.25$ for the fulfilled condition; $M = -1.76$, $SD = 2.03$) for the unfulfilled condition, $F(1, 86) = 0.07$, $p = .78$.

In this initial study, we found that, in three scenarios (i.e., donation, help, and hit), simply having intentions but not acting out on it did not indicate the morality of one's character to the same extent that acting out on such intention did. However, the scenario in which an agent showed intentions to leave a subway car due seeing a child with Down Syndrome showed a different pattern of results – the agent was rated as immoral regardless of the fulfillment of their intentions. Possibly, the nature of the transgressions may account for the different judgments for the prejudice scenario. The "donation", "help" and "hit" scenarios described individuals who intended to directly provide help or harm another, similar to what the Moral Foundations Theory would identify to violations of the "care" foundation (Haidt, 2001; Graham et al., 2013). On the other hand, the harm presented in the "prejudice" scenario is less straightforward, and seems more fitting to the purity foundation, related to moral transgressions associated with the body and associated with perception of contamination and feelings of disgust. To achieve more interpretable results, in Study 2 we modified the scenarios so that each of them pertained to either a *care* or a *fairness* moral concern (Haidt, 2007), based on the centric role of harm and justice for morality (Gray et al., 2012; Piazza et al., 2018).

The results found for the three consistent scenarios (i.e., donation, help, and hit) show that unfulfilled intentions do not lead to inferences of moral character as much as fulfilled intentions. However, because we did not provide an explanation to why the agent did not act out on their intention, it is possible that participants may have perceived the intentions as feasible or fleeting. Hence, in Study 2.2, we added a third condition of unfulfilled intentions, with the addition of an external explanation to its unfulfillment (i.e., unfulfilled-explained condition). With this condition, we expected that the external explanation provided would imply that the intention still remained, despite the agent's failure to act. To the extent that intention by itself influences the moral character evaluations, we hypothesized the actors in the "unfulfilled-explained" condition

would be rated as similar to the ones in the “fulfilled” condition (i.e., as moral as the latter for the positive vignettes and as immoral for the negative vignettes).

Study 2.2

Method

Participants

One-hundred and forty-two Japanese undergraduate students (72 male, M_{age} = 18.52, *SD* = 0.65) from a psychology class participated in the study in exchange for course credits. As explained in detail below, six vignettes were nested across the participants, and we expected that 20 participants were needed for each vignette based on the typical practice in the field, hence preparing for a total of 120 participants in minimum. We determined this sample size before any data analysis, and there was no exclusion of participants.

Materials, design, and procedures

Each participant read three positive vignettes depicting desirable behavioral intentions and three negative vignettes with undesirable intentions. Within each set of vignettes, we manipulated the fulfilment of the intention, with each vignette associated with either an unfulfilled-unexplained, an unfulfilled-explained, or a fulfilled act, as described in the Introduction section. Hence, we implemented a 2 (valence: positive vs. negative) x 3 (fulfillment: unfulfilled-unexplained vs. unfulfilled-explained vs. fulfilled) x 3 (vignettes) fully crossed, within-participant factorial design.

We aimed to expose each participant to all six vignettes, representing all conditions in the repeated measures design. Therefore, we created six different scenarios and nested them across the experimental conditions, using a Latin Square method (see Appendix 2 for full details). Specifically, the vignettes depicted six protagonists who intended to: (1) donate/steal money, (2) help/obstruct a classmate's academic activity, and (3) offer a help/force a burden to an elderly person (see Appendices B and C). We conducted a pilot study (*N* = 32) to ascertain that the three positive intentions were deemed to be positive, if fulfilled, using "morality" and "social

desirability" ratings (M = 3.94, SD = 0.77) on an 11-point scale (-5 = *extremely immoral/undesirable*, 5 = *extremely moral/desirable*). Ratings of the two items were combined because of a high correlation, rs > .62, ps < .001. Likewise, the three negative intentions were considered to be immoral and undesirable (M = -3.67, SD = 0.84).

In the main experiment, we presented the six vignettes to each participant in a randomly counterbalanced order, with the randomization controlled by the Qualtrics programming. After reading each vignette in the online platform, participants answered the following items concerning the dependent variable and potential mediators (all on 7-point scales).

Moral character. Participants rated how bad-good and immoral-moral the character was (-3 = *extremely immoral/bad*, 3 = *extremely moral/good*). We combined these items in our analyses of character evaluation, $r_s(140) > .77$, $p_s < .001$.

Emotions. Participants indicated how much positive (admiration and attraction, $r_s(140) > .61$, $p_s < .001$) and negative emotions (anger and disgust, $r_s(140) > .64$, $p_s < .001$) they felt towards the agent (1 = *no emotions at all*, 7 = *extreme emotio*ns).

Informativeness. We measured consistency and consensus as an index of informativeness, drawing on the rationale that the more frequently a person engages in a behavior (high consistency) and the less frequently other people behave similarly (low consensus), the more such behavior is diagnostic of disposition (Kelley, 1967). Thus, we assessed consistency by having participants rate the extent to which they believed the individual would repeat that thought/act (1 = *no repetition*, 7 = *certainly there will be repetition*). For consensus, participants indicated the extent to which they believed other people would have that thought/performed that act (1 = *no one at all*; 7 = *certainly many people*).

Results

We combined ratings on the same measure across the three positive and three negative vignettes, respectively, due to their consistent result pattern. Then, we conducted 2 (valence) x 3 (fulfillment) repeated measures Analyses of Variance (ANOVAs) on the morality, emotions, and informativeness scores. Table 6 shows the means and standard deviations for all measures.

Table 6. Means and Standard Deviations for Dependent Variables

		Positive vignettes		Negative vignettes	
Dependent variables	Fulfillment	*M*	*SD*	*M*	*SD*
Moral character	Unfulfilled unexplained	0.57	0.89	0.42	1.17
	Unfulfilled explained	1.24	1.08	1.76	0.96
	Fulfilled	2.25	0.98	2.14	0.88
Emotions	Unfulfilled unexplained	1.35	1.89	0.87	2.33
	Unfulfilled explained	2.74	1.95	3.04	2.16
	Fulfilled	4.13	2.08	3.38	2.21
Informativeness	Unfulfilled unexplained	0.51	1.71	0.49	1.57
	Unfulfilled explained	1.25	1.95	1.7	1.74
	Fulfilled	2.59	1.55	2.32	1.78

Moral character

The moral character scores were overall positive for the positive vignettes and negative for the negative vignettes. Hence, we reversed the scores of the negative vignettes to facilitate the comparison between valence. Higher scores represented greater perception of morality for the positive vignettes and greater perception of immorality for the negative vignettes. A 2 x 3 ANOVA revealed there was no significant main effect of valence, $F(1, 141) = 1.5$, $p = .22$, $\eta_p^2 = 0.01$, yet there was a significant main effect of fulfillment, $F(2, 282) = 239.27$, $p < .001$, $\eta_p^2 = 0.63$ with a significant two-way interaction, $F(2, 282) = 13.57$, $p < .001$, $\eta_p^2 = 0.09$. As Figure 9 illustrates, this interaction took place due to the asymmetric effect of the unfulfilled-explained

intention condition across valence. Specifically, the difference between the unfulfilled-explained and fulfilled conditions was smaller for negative vignettes ($M_{explained}$ = 1.76, SD = 0.96; $M_{fulfilled}$ = 2.14, SD = 0.88, CI 95% [0.19, 0.57]), $t(141) = 3.89$, $p < .001$, $d = 0.33$, compared to positive vignettes ($M_{explained}$ = 1.24, SD = 1.08; $M_{fulfilled}$ = 2.25, SD = 0.98), $t(141) = -10.14$, $p < .001$, d = 0.85, CI 95% [0.82, 1.22].

On the other hand, the difference between unfulfilled-unexplained and unfulfilled-explained conditions was larger for negative vignettes ($M_{unexplained}$ = 0.42, SD = 1.17; $M_{explained}$ = 1.76, SD = 0.96), $t(141) = 12.56$, $p < .001$, $d = 1.01$, CI 95% [1.13, 1.55], compared to positive vignettes ($M_{unexplained}$ = 0.57, SD = 0.89; $M_{explained}$ = 1.24, SD = 1.08), $t(141) = 5.98$, $p < .001$, d = 0.50, CI 95% [0.45, 0.89]. As hypothesized, intentions implied in the "explained" (rather than the "unexplained") condition led to higher evaluations of (im)moral character even when the acts were not fulfilled. This was especially the case when the intended acts were negative.

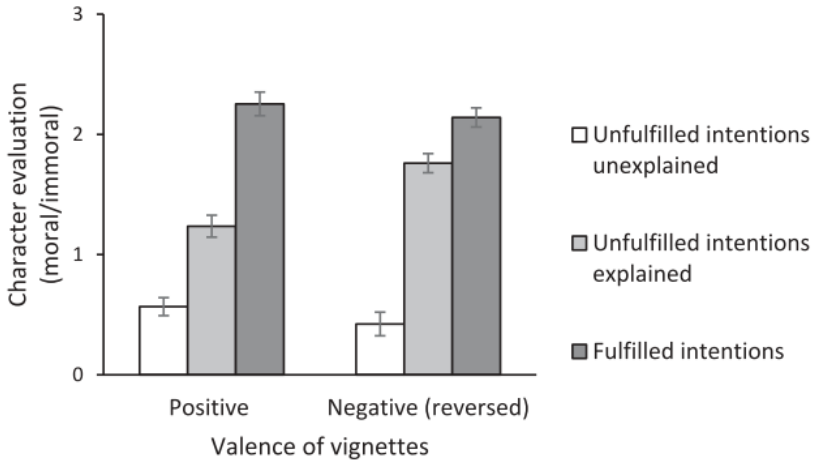

Figure 9. Moral character ratings across valence and fulfillment conditions. We observed a significant two-way interaction. The difference between the unfulfilled-explained and fulfilled conditions was smaller for the negative (vs. positive) vignettes. Bars represent standard errors. Figure adopted from Hirozawa et al. (2020).

Emotions

To construct a single measure of moral emotions, we subtracted "negative emotions" scores from "positive emotions" scores. Therefore, positive scores represented overall positive emotions towards the agent, whereas negative scores represented overall negative emotions. Similar to the morality results, participants felt overall positive emotions towards the agent in the positive vignettes and negative emotions in the negative vignettes. To make a more straightforward comparison concerning the net strength of the emotions between the valence conditions, we reversed the scores of the negative vignettes condition.

As Figure 10 shows, the positive-negative asymmetry in the effect of intention was even more evident. An ANOVA showed a marginally significant main effect of valence, $F(1, 141) = 3.57$, $p = .06$, $\eta_p^2 = 0.83$, and a significant main effect of fulfillment, $F(2, 282) = 148.25$, $p < .001$, $\eta_p^2 = 0.51$, qualified by a significant two-way interaction, $F(2, 282) = 7.04$, $p < .001$, $\eta_p^2 = 0.05$. Simple effect analyses showed participants reported similar negative feelings towards an agent who failed to fulfill their negative intention (with explanation) and an agent who actually fulfilled it ($M_{explained} = 3.04$, $SD = 2.16$; $M_{fulfilled} = 3.38$, $SD = 2.21$), $t(141) = 1.78$, $p = .08$, $d = 0.15$, CI 95% [0.04, 0.74]. In contrast, participants felt a significantly lower level of positive emotions towards an agent who did not fulfill their positive intention (vs. fulfilling), even when there was an external explanation ($M_{explained} = 2.74$, $SD = 1.95$; $M_{fulfilled} = 4.13$, $SD = 2.08$), $t(141) = 7.65$, $p < .001$, $d = 0.65$, CI 95% [1.03, 1.74].

Additionally, the difference between the unfulfilled-unexplained and the unfulfilled-explained condition was larger for negative vignettes ($M_{unexplained} = 0.87$, $SD = 2.33$; $M_{explained} = 3.04$, $SD = 2.16$), $t(141) = 10.37$, $p < .001$, $d = 0.87$, IC 95% [1.76, 2.58], compared with positive vignettes ($M_{unexplained} = 1.35$, $SD = 1.89$; $M_{explained} = 2.74$, $SD = 1.95$), $t(141) = 6.14$, $p < .001$, $d =$

0.51, CI 95% [0.95, 1.84]. In sum, the more an agent was perceived as intentional (implied by the explanation), the higher were the emotions evoked, particularly when the intentions were negative (vs. positive).

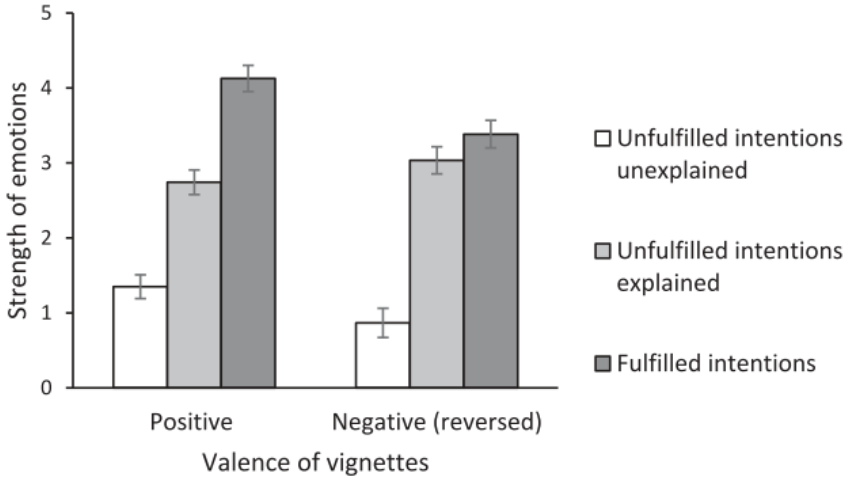

Figure 10. Ratings of strength of emotions across valence and fulfillment conditions. Again, there was a significant two-way interaction. The difference between the unfulfilled-explained and fulfilled conditions turned insignificant for the negative vignettes, but significant for the positive vignettes. Bars represent standard errors. Adopted from Hirozawa et al. (2020).

Again, there was a significant two-way interaction. The difference between the unfulfilled-explained and fulfilled conditions turned insignificant for the negative vignettes, but significant for the positive vignettes. Bars represent standard errors.

Informativeness

We measured the extent to which one's intentions/acts are informative of character through the "consistency" and "consensus" items. We reasoned that the more people perceive intentions/acts as consistent, the greater is its informativeness of character. On the other hand, the perceived informativeness of character should vary in proportion to the perceived level of consensus of these intentions/acts. According to the literature of causal judgments, these two elements of consistency and consensus relate in a multiplicative way (Cheng & Novick, 1990;

Forsterling, 1989; Kelley, 1967). Therefore, we divided the "consistency" scores by the "consensus" scores to create a single measure of informativeness.

The analyses revealed no significant main effect of valence, $F(1, 141) = 0.25$, $p = .62$, $\eta_p^2 = 0.00$, yet a significant main effect of fulfillment, $F(2, 282) = 46.44$, $p < .001$, $\eta_p^2 = 0.25$. The two-way interaction was not significant, $F(2, 282) = 0.79$, $p = .46$, $\eta_p^2 = 0.01$. Therefore, differently from the asymmetric results for our other measures, positive and negative intentions showed no difference in the extent to which they diagnosed character.

For both positive and negative intentions, failing to fulfill one's intentions with no explanation led to significantly weaker inferences of (im)moral character compared to the condition with the explanation ($M_{unexplained} = 1.29$, $SD = 0.61$; $M_{explained} = 1.65$, $SD = 0.73$), $t(141) = -4.45$, $p < .001$, $d = 0.38$, CI 95% [-0.52, -0.20]. Likewise, the mean informativeness score in the unfulfilled-explained condition was also significantly weaker compared to the fulfilled condition, ($M_{explained} = 1.65$, $SD = 0.73$; $M_{fulfilled} = 2.09$, $SD = 0.88$), $t(141) = -5.46$, $p < .001$, $d = 0.46$, CI 95% [-0.60, -0.28].

Mediation

One main purpose of this study was to investigate the psychological processes (i.e., emotions and informativeness) underlying the predicted effect of unfulfilled intentions on judgments of moral character. To attain this goal, we conducted multiple mediation analyses. As our measures were repeated within participants, we performed Bootstrapping analyses (5,000 resampling) using the MEMORE macro for SPSS (Montoya & Hayes, 2017).

In the MEMORE macro for repeated measures, the effect of the independent variable is represented by the subtraction of the two repeated-measures observations. In our study, we assumed the vignettes of the unfulfilled-unexplained condition implied that the actor had

plausibly lost their intention, whereas we expected the vignettes of the unfulfilled-explained condition to suggest that the intentions remained regardless of their fulfillment. Hence, we reasoned that the difference between the scores of the unfulfilled-explained and unfulfilled-unexplained conditions demonstrated the effect of the intention implied by the explanation manipulation. Hence, the effect of intentions on moral character evaluations was represented by the difference between these conditions for moral character scores, i.e., unfulfilled-unexplained subtracted from unfulfilled-explained (Moral Caracter$_{explained}$ – Moral Character$_{unexplained}$). Moreover, we tested whether the effect of this difference on moral character was mediated by two hypothesized mediators, namely emotions and informativeness. Our first mediator was coded as the difference between these conditions for the emotion scores (Emotion$_{explained}$ – Emotion$_{unexplained}$). Finally, the second mediator was coded as the difference between these conditions for the informativeness scores (Informativeness$_{explained}$ – Informativeness$_{unexplained}$). We conducted these mediational analyses separately for the positive and negative cases (see Figure 11).

For the positive vignettes, emotions showed a significant indirect effect (*ab* = 0.52, 95% CI = [0.36, 0.69]), while the mediation by informativeness was not significant (*ab* = -0.01, 95% CI = [-0.05, 0.04]). The direct effect of intention was not significant when taking the mediators into account (*c'* = 0.15, *p* = .074). The total effect was *c* = 0.67, *p* < .001, 95% CI = [0.45, 0.89].

In contrast, for the negative vignettes, the indirect effects of emotions (*ab* = 0.74, 95% CI = [0.55, 0.96]) and informativeness (*ab* = 0.06, 95% CI = [0.01, 0.12]) were both significant. The direct effect of the independent variable (i.e., intention) remained significant after we took the mediators into account (*c'* = 0.54, *p* < .001, 95% IC = [0.35, 0.74]). The total effect was *c* = 1.34, *p* < .001, 95% CI = [1.32, 1.36]

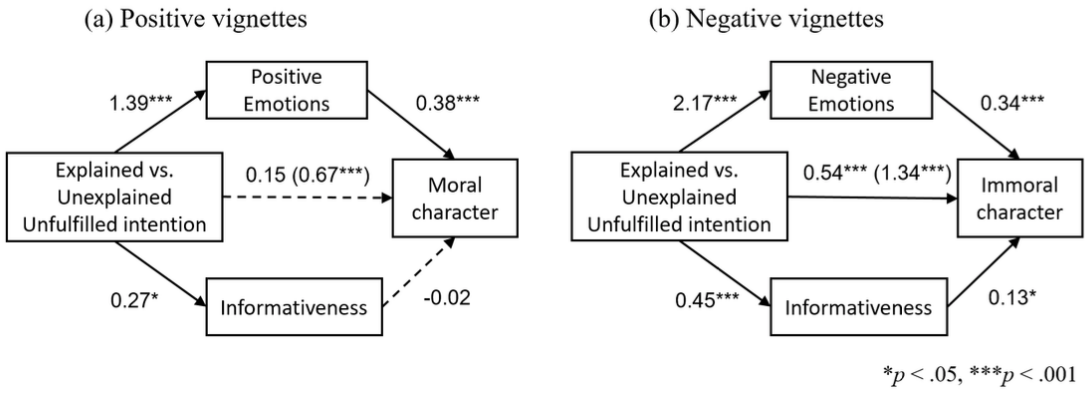

Figure 11. Mediation analysis using MEMORE for positive and negative vignettes. Intention implied by the justification manipulation was coded as the difference between the unfulfilled-explained condition and the unfulfilled-unexplained condition. For the positive vignettes (a), the effect of intentions on moral character was mediated by emotions but not informativeness. For the negative vignettes (b), the effect of intentions on moral character was mediated by both emotions and informativeness. Adopted and adapted from Hirozawa et al. (2020).

General Discussion for Study 2

In two studies, unfulfilled intentions informed little of character compared to fulfilled intentions, possibly because there was no explanation to why the actor behaved inconsistently with their intentions, which may have indicated that the intentions were fleeting. In Study 2.2, we included an important condition: the unfulfilled intentions with an explanation, so to ascertain that participant would infer the presence of the intention. The results revealed that greater intentions to perform positive and negative acts led to higher ratings of moral and immoral character. This effect took place despite no factual action followed the mental state (i.e., unfulfilled and unfulfilled-explained conditions). Importantly, we found that intentions affected moral judgments in an asymmetric way, that is, negative intentions had a greater impact on consequent moral character evaluations, compared to positive intentions. The mediational analyses suggesting the underlying psychological processes also revealed an asymmetry towards a negativity bias. For the positive scenarios, the link between intentions and moral character was fully accounted for by emotions. In contrast, for the negative scenarios, both emotions and informativeness were significant mediators. Hence, the informativeness of positive intentions did not necessarily lead to the judgment of one's morality, and inferences of character were driven mainly by positive emotions felt towards the agent. On the other hand, negative intentions had a greater weight on inferences of moral character, explained by both the negative emotions felt towards the agent, and by the extent to which such intentions were perceived as more frequent and less consensual.

Our results support the relevance of intentions on moral judgments (Ames, & Fiske, 2013; Cushman, 2008). Previous research mostly explored the effect of intentions on moral evaluations after the act was completed. To our knowledge, this is the first study to directly show the influence of intentions that *were not acted out* on moral character judgments. We also extend the

current literature by examining the asymmetric effect of intentions. Our results were in line with the vast literature concerning greater weight of negative (versus positive) information on various psychological processes (see Baumeister et al., 2001).

In searching for explanations to the effect of intentions on moral character evaluations, it is noteworthy that emotions turned as significant mediators for both positive and negative vignettes. Our results suggest that people not only feel good about good-doers and bad about bad-doers, but they also feel good about well-intentioned people and bad about ill-intentioned people on the basis of their intention alone. From a consequentialist perspective, this sounds illogical. However, this automatic, emotion-based response to mental state is likely how people are wired from early on in the development process. For instance, classic developmental research shows that children display an increasing consideration of intentions (versus outcomes) in moral judgments as they grow (Piaget, 1965). Even more suggestive of an early automatic process is the finding that infants as young as eight-months-old showed sensitiveness to mental states, preferring puppets who tried to help (vs. prevent) a third party achieve its goal, regardless of succeeding or not (Hamlin, 2013).

Furthermore, as especially negative intentions increased evaluations of immorality of character in the absence of harm and even action, our results contrast with harm-centered approaches of morality (Gray et al, 2012), which posit harm as a necessary aspect of moral judgment. However, perceived intention may also have driven inferences of character by an underlying perception of potential harm. That is, it is sensible that an individual with negative intentions seems to represent a potential harm to society, and hence is considered to be immoral. Our results on informativeness support this idea, as negative intentions were revealing of immoral character partly because they were perceived as consistent and low in consensus, hence more likely to occur again and generate harm. Future studies may investigate this possibility

more thoroughly. If so, this account would also align with the person-based approach of morality, which theorizes that individuals invest in predicting immoral characters as an advantageous strategy to avoid the potential risks these individuals represent (Pizarro, & Tannenbaum, 2011).

The results on informativeness, or lack thereof, from the positive side were also noteworthy. Even if one's positive intentions were perceived as frequent and distinct, they were still not diagnostic of moral character. This is puzzling finding, as it contradicts frequency-based explanations of diagnosticity (Mende-Siedlecki et al., 2013). We also notice that this asymmetry was found even though the informativeness scores showed no difference between positive and negative vignettes (i.e., insignificant Valence x Condition interaction). A possible explanation for this result is that people intuitively assume that others should be good-intentioned by default. In this sense, even if one has distinctively more good intentions than others, it still corresponds to the desired social norms and hence does not particularly reveal one's character. Supportive of this idea, Pizarro et al. (2003) found participants did not expect an agent who emotionally broke someone's car window to have desired to do so, that is, they did not expect one to possess negative desires. Other studies suggest people are intuitively prosocial. For instance, Rand and Epstein (2014) found the decision-making process of altruists who risked their own life to save others is highly intuitive (vs. deliberate). Another study showed participants made more prosocial decisions in an economic game when under cognitive load (Rand et al., 2014). In sum, it is possible that positive intentions, even if frequent and distinct, are not as diagnostic of one's moral character as negative intentions are of immoral character because they are overall predicted to be the default mental state, and therefore require stronger evidence (such as acting out) to indicate one's moral disposition.

Finally, our study includes several limitations. First, our manipulation of intention consisted of an inference of intention based on the external explanation. Future studies may

explore different measures of intentions. In addition, our sample consisted of exclusively Japanese university students, which may hinder the generalizability of our findings. However, we also note that, compared to Westerners, East Asians are more likely to resist correspondence bias by considering the broad social context (Choi et al., 1999). As we found a clear effect of unfulfilled intentions on moral character even for this population, we speculate our results may represent a more fundamental, universal psychological process.

Our study was also limited to the asymmetric effect of unfulfilled intention on inferences of character. Future studies shall explore whether this effect of intentions extends to behavioral consequences towards the (im)moral agent, such praise/blame and reward/punishment. For instance, in criminal law, we punish attempted crimes with no concrete fulfillment nor outcome, such as conspiracies (Christopher, 2004). It is also not hard to imagine ordinary situations in which people make decisions based on underlying perceived mental states. People break friendships at knowing their ex-friend's prejudiced beliefs, even if the friend would not display discriminative behavior. Married ones sign for divorce after finding out their partner contemplated intentions of cheating. Civilians want to punish politicians based on their proposals, regardless of actual implementation. In this study and from a person-based approach to moral judgments (Pizarro, & Tannenbaum, 2011), a possible reason why people punish or avoid evil mental states is because of their primary motivation to punish and eliminate a bad person from society. A second possibility, as we have discussed, is the underlying perception of potential harm. Further research might uncover the psychological processes underlying such judgments.

Summary

This study investigated whether intentions by themselves are predictive of an agent's moral character. The answer found was yes, but with a detail. Both positive and negative intentions elicited inferences of moral and immoral character, respectively, due to emotional

reactions elicited by the intention information. However, only negative intentions seemed particularly informative of character (i.e., significant mediation by informativeness). Positive intentions are likely perceived to be the default mental states, and its consistency/consensus ratio did not predict moral character.

Chapter 4 - Study 3. Mental states effect on blame for failure to help

In Study 1, empirical evidence demonstrated that people who choose to omit help (vs. take action to help) were blameworthy in part because their actions were informative of an underlying immoral character. In the present chapter, we extend this finding to another important element of blame: mental states. Specifically, this study investigates whether individuals also blame failure to help based on the extent to which the agent's mental states are indicative of their immoral character. Study 2 provided evidence that positive and negative intentions in fact do drive inferences of both moral and immoral character, respectively, but especially so for the latter. In this study, we also searched for the differential effects that the valence of mental states may produce in blame.

Specifically, in this study we reduced the conceptualization of mental states to desires. Desires reflect what individuals wish and want, and desires by themselves do not imply action. For instance, a person may want to save the world but that is different from deciding to act on it (Malle & Knobe, 1997). If information on one's positive and negative desires affect judgments of blame, then there's a clearer evidence that mental states alone influenced ascriptions of blame, irrespective of implied action.

This study involved scenarios of agents who promised to help another person yet unintentionally failed to help (e.g., Nakamura promises he would take his friend to the airport. However, Nakamura ends up unable to keep his promise due to an urgent last-minute meeting at work and, as a consequence, his friend misses his flight). We manipulated the extent the agent desired to help. In the "negative desires" condition, the agent does not desire to help and has negative wishes towards the promised individual. On the other hand, in the "positive desires" condition, the agent desires to help and has positive wishes towards the promised individual.

Finally, there was a "neutral desire" condition, in which the agent does not particularly want to help but does not oppose it either. Considering the positive-negative asymmetry found in the literature, we expected the agent with negative (vs. positive) desires to receive the greatest level of blame. Establishing a neutral condition also allowed us to examine whether information of positive desires mitigates blame.

This prediction was based on evidence that desires seem to affect blame regardless of associated intentionality or causality perceptions. In Inbar et al.'s (2012) set of four studies, they found that individuals who benefited from a misfortune (e.g., winning a bet that a natural disaster will occur) were deemed more blameworthy for their acts compared with those who did not benefit from it, despite having no causal or intentional role in the disaster occurrence. The results were explained by underlying perceptions of negative, or "wicked" desires. Based on the person-centered approach to moral judgments (Pizarro &Tannenbaum, 2011; Uhlmann et al., 2015), people are motivated to blame immoral agents, and hence benefitting from misfortune received greater blame because of an underlying assumption that only bad agents would have such wicked desires.

In Inbar et al.'s (2012) study, blameworthiness was measured by a combined composite of blame and wrongness. However, Cushman (2008) suggests that these two judgments are distinct. Judgments of wrongness are more sensitive to mental states information, such as beliefs and desires, whereas blame is determined by both mental state and outcome information. Helping others and fulfilling one's social obligations are commendatory and normative behaviors (Janoff-Bulman et al., 2009) that are socially reinforced (Graham et al., 2013; Shweder et al., 1997). It seems people expect others not only to behave positively but to also present corresponding positive mental states. For instance, people seem to infer that individuals in general have good desires (Pizarro et al., 2003), and good intentions are not as diagnostic of character compared to

negative intentions (Hirozawa et al., 2019), suggesting that positive mental states are deemed as default. In general, these findings suggest that people expect others to act with positive desires, which implicates that acting with negative desires may be perceived to be a wrongful course of action. To examine this possibility, we treated wrongness as a potential mediator of the effect of desires on blame.

As previously noted, a second mechanism through which people may assign greater blame for failure to help with negative desires is based on person evaluations. The person-centered approach to moral judgments posits that blame is assigned with the motivation to blame immoral characters inferred by their mental state (Pizarro &Tannenbaum, 2011). Previous research shows evidence of a dissociation between moral judgments of acts and that of persons. For instance, bigots received greater blame compared to physical assailants, even though racial slur was perceived to be a less immoral violation than physical assault (Uhlmann et al., 2013). People are deemed blameworthy even when performing harmless acts, as long as their behavior seems informative of a bad character (see Pizarro et al., 2012). Hence, people may determine blame for failure to help not only on the basis of the wrongness of the act but also on what the act informs about the person. We also examined the potential role of perceived moral character as a second mediator.

Finally, we elaborated the scenarios in a way that the perceptions of causality and intentionality should remain constant. That is, regardless of the desire condition, all agents failed to keep their promise unintentionally and for the same reasons. However, a motivated reasoning account suggests that mental state information can alter following factual moral judgments. For instance, an agent was perceived as more causal of an accident when he was driving home to hide cocaine (vs. hide a gift for his parents) (Alicke, 1992). People also assigned greater harm to intentional (vs. unintentional) acts even when the harm was identical (Ames & Fiske, 2013). This

account predicts that people may inflate moral judgments to justify their blame motivation. To address to this potential explanation, we also included measures of causality and intentionality.

In sum, in the present study, we examined whether individuals who unintentionally fail to help with negative desires (vs. positive and neutral desires) receive greater blame. We hypothesized judgments of wrongness of their failure to help and inferences of immoral character to be potential mediators of this effect. To account for a potential blame validation phenomenon, we also included causality and intentionality in the mediational analyses.

Method

Participants

We recruited 210 participants through an outsourcing service, CrowdWorks (112 females, M_{age} = 38.78, *SD* = 8.77). The sample size was determined by prior analysis using GPower. Due to the nature of the Latin Square design, we chose for a conservative calculation based on each pair of scenarios, that is, a between-subjects Analysis of Variance for main effects and interaction (η_p^2 = 0.09, α = .05, 1 - β = .80, *df* = 2, *groups* = 6). The minimum sample size for a set of two scenarios was 101. We determined the sample size prior to any data analysis, and there was no exclusion of participants.

Materials, design, and procedures

We prepared six scenarios describing agents who failed to keep their promise to help another person, followed by a negative outcome. Specifically, the scenarios depicted agents who failed to keep the promise to: (1) donate to their friend's project, which ended up getting cancelled for lack of donations; (2) buy a birthday cake for their friend, spoiling the birthday party; (3) help their cousin with moving out, resulting in the cousin having to pay a fee to the real state agency; (4) help a classmate with math, resulting in the classmate's failure in the test; (5) take care of their

sick grandfather, who became more ill; (6) take a friend to the airport, who ended up missing their flight. In all scenarios, the agents made the promise compelled by either the promised person or a third person (e.g., the friend asked for a lift to the airport; the agent's mother asked him to help his cousin). The agents failed to help due to forgetfulness or an external justification (e.g., they realized they had no money when they were about to donate). We separated these scenarios into one set (A) containing scenarios 1, 2, and 3, and another set (B) containing scenarios 4, 5, and 6. Participants were randomly assigned to either one of these sets (see Appendix 3 for examples of full scenarios).

Moreover, as detailed in the introduction, we manipulated three levels of desires. For example, in the Airport scenario, the agent's desires were described as either:

> **Positive.** Deep inside, Hirata really wanted to drive his friend to the airport. He thought friends should help each other and wished his friend would have a pleasant journey.
>
> **Neutral.** Deep inside, Hirata didn't particularly want to drive his friend to the airport but didn't mind doing it either. He felt neutral about his friend's request.
>
> **Negative.** Deep inside, Hirata really did not want to drive his friend to the airport. He thought his friend was inconvenient and wished his friend would have a hard time in his journey.

We rotated the three levels of the desire manipulation across the scenarios using a Latin Square method (see Table 7 for details). Each participant was presented to either the set of scenarios A or B and went through all levels of the desire manipulation. The presentation of the scenarios was randomly counterbalanced. Hence, we implemented a 2 (Set: A vs. B) x 3 (Desires: positive vs. neutral vs. negative) x 3 (Scenarios) fully crossed, within-participant factorial design, and with Set as a between-subjects variable.

Table 7. Latin Square Design Arrangement

Latin Square Design Arrangement						
Set A						
A1						
	1	Positive	2	Neutral	3	Negative
A2						
	1	Negative	2	Positive	3	Neutral
A3						
	1	Neutral	2	Negative	3	Positive
Set B						
B1						
	4	Positive	5	Neutral	6	Negative
B2						
	4	Negative	5	Positive	6	Neutral
B3						
	4	Neutral	5	Negative	6	Positive

Note: Numbers "1" to "6" stand for the previously listed scenarios. Each participant went through either the three combinations from set A (i.e., A1, A2, and A3) or the three combinations from set B (i.e., B1, B2, and B3).

After reading each scenario in the online platform, participants answered on 7-point scales:

Valence of desires. The first item of desires read, "In the scenario, it is informed [the agent]'s desires and thoughts about helping. To which extent do you think [the agent]'s desires and thoughts were negative/positive?" (*1 = very negative, 7 = very positive*). Participants also rated the extent to which they thought the agent's desires and thoughts were desirable (*1 = very undesirable, 7 = very desirable*), $rs > .72$, $p < .001$.

Blame. Participants indicated how much they blamed the agent for failing to help (1 = *no blame at all*, 7 = *extreme blame*) and for the negative outcome that followed their failure to help (1 = *no blame at all*, 7 = *extreme blame*), $rs > 0.76$, $p < .001$.

Wrongness. Participants indicated the extent to which the agent's failure to keep the promise was wrong, (*1 = not wrong at all, 7 = extremely wrong*).

Immoral character. Participants rated how bad-good, immoral-moral, untrustworthy-trustworthy the character was (1 = *extremely moral/ good/ trustworthy, 7 = extremely immoral/ bad/ untrustworthy*). We combined these items in our analyses of immoral character, αs > .84.

Causality. Participants rated the extent to which the agent was the cause of the negative outcome (*1 = not the cause at all, 7 = definitely the cause*).

Intentionality. Participants rated the extent to which the agent's failure to help was intentional (*1 = not intentional at all, 7 = definitely intentional*).

Results

Due to our Latin Square design, it was only possible to analyze the intention manipulation as a between-subjects variable when focusing on each individual scenario. Analyses on each scenario for all measures suggested a consistent pattern of results for the main dependent variable, i.e., blame. Therefore, we reanalyzed the data with the intention manipulation as a within-subjects measure, hence combining all scenarios. We conducted One-way repeated measures Analyses of Variance (ANOVAs) on all of the dependent variables. Table 8 shows the means and standard deviations for all measures.

Manipulation check

There was a significant main effect of desires on the valence of desires scores ($F(2, 208) = 104.69$, $p < .001$, $\eta_p^2 = 0.34$), revealing the successfulness of our manipulation. Pairwise comparisons with Bonferroni correction revealed that participants rated the desires of the agent as more positive for the positive desires condition compared to the neutral ($p < .001$) and negative conditions ($p < .001$). The agents' desires in the neutral condition also received greater ratings compared to the negative desires condition ($p < .001$).

Blame

Consistent with our predictions, the more the desires of the agent were negative, the greater were the blame scores ($F(2, 416) = 5.12$, $p = .006$, $\eta_p^2 = 0.02$). Specifically, an agent with negative desires was more blameworthy than one with positive desires ($p = .03$). However, agents with negative and neutral desires received blame to a similar extent ($p = 1.00$). This result suggests that it does not take overt negative mental states to invite blame. Displaying indifference to helping others (i.e., neutral condition) was also judged as equally blameworthy. Finally, agents with positive desires were less blamed than those with neutral desires ($p = .005$).

Wrongness

The effect of desires on wrongness scores was only marginally significant ($F(2, 416) = 2.59$, $p = .08$, $\eta_p^2 = 0.01$). There was a tendency for harsher judgments of wrongness the more the desires were negative. However, at a $p < .05$ level of significance, the effect of desires on wrongness was not significant.

Immoral character

The more the desires were negative, the more the agent was perceived to be an immoral character ($F(2, 416) = 55.14$, $p < .001$, $\eta_p^2 = 0.21$). An agent with negative desires was evaluated as being more immoral than an agent with positive and neutral desires ($ps < .001$). The agent with neutral desires was also perceived as having poorer character compared to the one with positive desires ($p < .001$).

Causality

The effect of desires on the causality scores was non-significant, $F(2, 416) = 0.09$, $p < .91$), suggesting that ascription of causality was the same for all levels the desires manipulation.

Intentionality

Although all scenarios described the agent's failure to help as unintentional, there was a significant main effect of desires on intentionality, $F(2, 416) = 46.98$, $p < .001$, $\eta_p^2 = 0.19$). The more the desires were negative, the more the failure to help was perceived to be intentional. All pairs were significantly different from each other ($ps < .05$).

Table 8. Means and standard deviations for each dependent variable

Means and Standard Deviations (in parenthesis)			
	Valence of desires		
DVs	*Positive*	*Neutral*	*Negative*
Valence of desires	5.09 (1.55) a	4.13 (1.27) b	3.42 (1.21) c
Blame	3.95 (1.58) b	4.32 (1.51) a	4.28 (1.58) a
Wrongness	4.03 (1.63) a	4.25 (1.60) a	4.33 (1.66) a
Immoral character	3.75 (1.26) c	4.38 (1.10) b	4.75 (1.04) a
Causality	4.28 (1.85) a	4.30 (1.81) a	4.22 (1.86) a
Intentionality	2.00 (1.32) c	2.24 (1.28) b	3.03 (1.55) a

Note: The subscripts "a", "b", and "c" represent the pairwise differences, with higher means ordered in alphabet order. Cells sharing different subscripts in each row were significantly different from each other. Cells sharing the same subscripts were not significantly different from each other.

Mediation

To investigate the psychological processes underlying the effect of desires on blame, we conducted parallel multiple mediation analyses. We performed Bootstrapping analyses (10,000 resampling) using the MEMORE macro for SPSS (Montoya & Hayes, 2017), which is an

adequate tool for our repeated measures design, and which calculates the effect of the independent variable by the subtraction of the two repeated-measures observations.

We show three set of mediation analyses. In the first set, we compared the positive versus negative desires conditions for a larger perspective of how increased negative intentions affect blame (see Figure 12). The mediation analyses revealed that the increased negativity of the desires affected blame indirectly by increasing perceptions of wrongfulness (*ab* = 0.09, 95% CI = 0.01 to 0.18) and immoral character (*ab* = 0.29, 95% CI = 0.15 to 0.44). The indirect effects by causality (*ab* = -0.01, 95% CI = -0.13 to 0.09) and intentionality (*ab* = -0.01, 95% CI = -0.11 to 0.09) were non-significant. The direct effect was *c'* = 0.03, *p* = .70, 95% CI = -0.23 to 0.15. The total effect was *c* = 0.33, *p* < .001, 95% CI = 0.32 to 0.35.

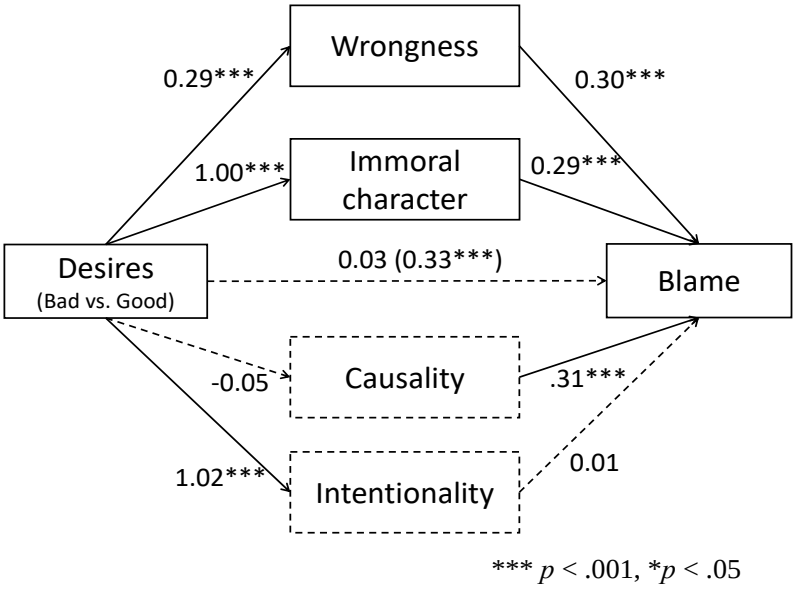

Figure 12. Parallel mediation analyses revealed that the effect of desires (bad vs. good) on blame for failure to help was mediated by evaluations of wrongness and immoral character. Causality and intentionality were not significant mediators.

In the second set of mediation analyses, we compared the bad and neutral desires conditions to reveal the sole role of the negativity of intentions on blame (see Figure 13). In this comparison,

increased negativity of the desires affected blame indirectly by increasing perceptions of immoral character (*ab* = 0.14, 95% CI = 0.05 to 0.26). The indirect effects by wrongness (*ab* = -0.02, 95% CI = -0.10 to 0.04), causality (*ab* = 0.02, 95% CI = -0.10 to 0.16) and intentionality (*ab* = 0.08, 95% CI = -0.01 to 0.18) were non-significant. The direct effect was *c'* = 0.08, *p* = .58, 95% CI = -0.08 to 0.25. The total effect was *c* = 0.03, *p* = .002, 95% CI = 0.01 to 0.04.

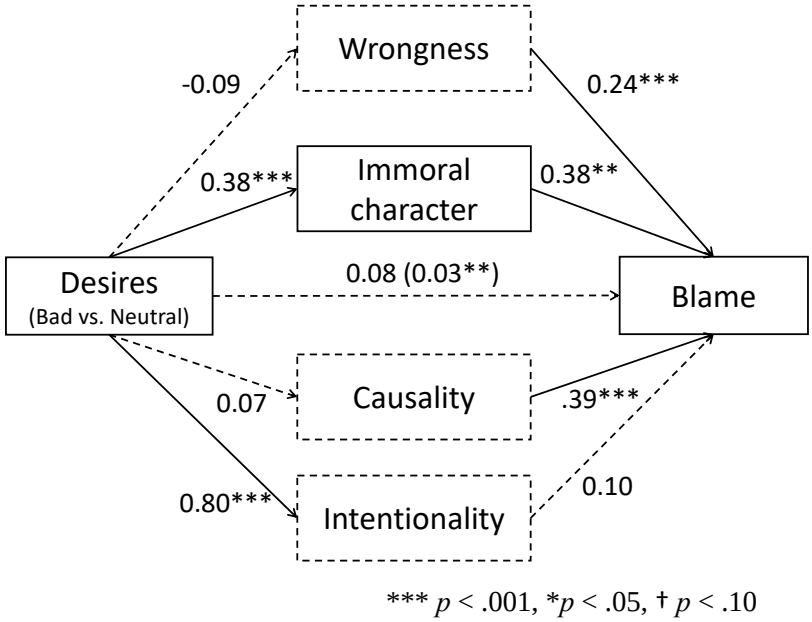

Figure 13. Parallel mediation analyses revealed that the effect of negative desires (bad vs. neutral) on blame for failure to help was mediated by evaluations of immoral character. Wrongness, causality and intentionality were not significant mediators.

Finally, in the third set of mediation analyses, we compared the positive and neutral desires conditions to reveal the role of the positivity of intentions on blame (see Figure 14). Increased positivity of the desires mitigated blame indirectly by increasing perceptions of moral character (*ab* = 0.13, 95% CI = 0.03 to 0.23). This measure of moral character corresponds to the reversed measure of immoral character ratings to facilitate the interpretation of the results. The indirect effects by wrongness (*ab* = -0.05, 95% CI = -0.01 to 0.12), causality (*ab* = -0.01, 95% CI = -0.11

to 0.10) and intentionality (*ab* = -0.03, 95% CI = -0.06 to 0.01) were non-significant. The direct effect was *c'* = -0.15, *p* = .07, 95% CI = -0.01 to 0.31. The total effect was *c* = 0.37, p < .001, 95% CI = 0.35 to 0.38.

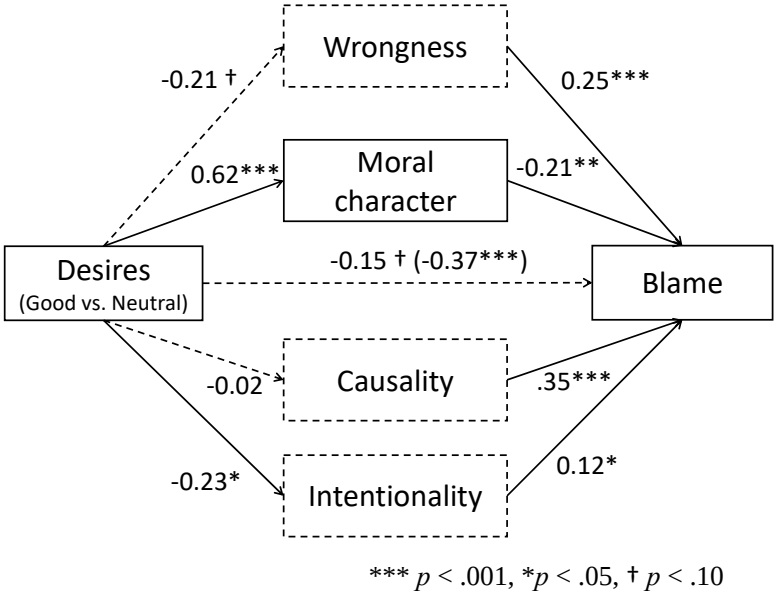

Figure 14. Parallel mediation analyses revealed that the mitigating effect of positive desires (good vs. neutral) on blame for failure to help was mediated by evaluations of moral character. Wrongness, causality and intentionality were not significant mediators.

Discussion

In this study, agents who unintentionally failed to help another person were considered to be more blameworthy when their desires were negative (vs. positive). Curiously, our findings also revealed that agents with neutral desires, that is, who did not particularly desire to help but did not oppose it either, received similar blame judgments to those who openly carried negative desires. As discussed in the introduction, this supports the premise that individuals expect others to not only "not have" bad desires, but to actually display positive ones.

The significant mediation by wrongness in the effect of negative desires on blame demonstrates a normative aspect of blaming failure to help. Individuals are not expected to help

with negative desires, and doing so increases the assignment of blame in case one fails to help. These findings suggest that wicked desires increase blame in part because they are perceived to be wrong. We note that this role of wrongness should be sensitive to individual and cultural differences. For instance, Protestants (vs. Jews) endorse to a greater extent in beliefs about the moral relevance of thoughts, which is associated with a greater role of mental states in person evaluation for this group (Cohen & Rozin, 2001). In cultures which avoid mental state reasoning, mental states are less relevant for moral judgments (McNamara et al., 2019).

The three set of mediation analyses also revealed that the effect of the agent's mental states on blame was explained by inferences of character. The more negative were the desires, the more participants inferred the agent was immoral, and blame judgments increased. On the other hand, the more positive were the desires, the more participants mitigated blame. These results cast further light on the apparent illogical role of desires on blame. To the extent that the effect of desires did not affect blame through perceptions of causality and intentionality, rationalist theories of blame would predict that this information would not be relevant in determining blame (see Malle et al., 2014). However, according to the person-centered approach to moral judgments, blame is assigned based on inferences of character underlying one's actions and, in this case, one's desires. Our findings corroborate with this theory and replicate Inbar et al.'s (2012) results with a direct mediation test.

Finally, consistent with the literature on blame, causality judgments predicted blame. However, increased negativity of the desires did not predict inflated judgments of causality, leading to greater blame (i.e., the mediation by causality was non-significant). The literature on motivated reasoning suggests that individuals who are motivated to blame may inflate factual moral judgments, such as causal attribution, to support such motivation (Alicke, 1992; Alicke, 2000). In this study, we did not find evidence of this process. Likewise, more negative desires

were perceived as more intentional, consistent with the premise that desires are a fundamental component of intentionality (Malle & Knobe, 1997). Nevertheless, intentionality did not explain the effect of desires on blame. Overall, these findings rule out the alternative explanation that the mediational roles of wrongness and moral character were a byproduct of underlying perception of intentionality and cause.

An important limitation of this study is that we focused our scenarios only on ordinary examples of unintentional failure to help. Perhaps because of the subtle level of harm in these scenarios, the effect of desires on blame was small. An open question is whether the same processes are replicated in judgments of criminal cases, such as omissions. For instance, this study could be extended to cases of medical negligence (e.g., a doctor that had no desire to treat a patient and unintentionally fails to treat them) or parental negligence (e.g., parents who do not desire to care for their child and unintentionally fail to take care of them). Exploring these possibilities may show clearer evidence of the blame processes found in this study. As previously discussed, our findings may also be highly sensitive to cultural influence. Future studies may examine whether these findings will replicate across cultures.

To our knowledge, this is the first study to investigate the role of desires on blame in the context of unintentional failure to help. This study has important implications for the research on moral judgments and justice. We revealed that, in laypersons' judgments, desires determine blame for failure to help to the extent that they signal deviation from the norm and inform immoral character. However, from a legal perspective, desires that do not inform intentionality should be irrelevant for judging blame, and individuals should be punished on the basis of their actions and not their character (Dressler, 2015). This suggests potential biases in the process of judging cases of omission. Further studies may directly explore the processes of blame for failure to help in cases of omission.

Summary

This study demonstrated that agents who failed to help with negative (vs. positive) desires were more deemed more blameworthy due to underlying perceptions of immoral character. This finding took place even after taking into account perceptions of causality and intentionality. Moreover, the more intentions were positive, the more blame was discounted. This study contributes to the literature by showing empirical evidence of the role of moral character inferences on judgments of blame for failure to help.

Chapter 5 – Conclusion

The commandment that one should help others is present across the moral foundations of diverse complex societies (Barret et al., 2002; Boster et al., 2001; Curry et al, 2019). To the extent that one has a social duty to help, it follows that failure to do so should evoke blame. In such situations, how do people ascribe blame for an agent who fails to help? The present thesis investigated the psychological processes underlying blame for failure to help. Blame has been mainly approached from the perspective of proscriptive immoralities. That is, there is a great deal of studies which investigates the psychological process of blame for committed negative acts (Alicke, 1992; Cushman, 2008; Malle et al., 2014; Spranca et al., 1990; Wells & Gavenski, 1989). Therefore, it is of theoretical relevance to explore whether the same principles that determine judgments of blame for proscriptive immoralities apply to prescriptive immoralities.

There are two important elements for judgments of the blameworthiness of an agent in both lay person's judgment and in criminal law. The first is the agent's guilty action or inaction. People are blamed by both committing negative actions or by omitting positive action. The second element concerns the agent's guilty mind. As discussed in Chapter 1, human beings tend to assume that other individuals have a mind, and they use such inferred mental states (e.g., intentions, thoughts, beliefs) to predict behavior and make judgments. Finally, the person-based approach of moral judgments suggest that individuals blame others based on assessments of moral character (Uhlmann et al., 2014). In this thesis, I explored the psychological processes underlying the effect of action-inaction and mental states on blame for failure to help, focusing on the distinct role of moral character.

In Study 1, I tested the effect of failing to help by action (i.e., taking action to help yet failing; commission) versus inaction (i.e., not taking action to help; omission) on blame for

failure to help and investigated the role of moral character as a potential mediator of this effect. The results showed that omitting help (vs. taking action) elicited greater blame in part because an agent who omitted help was perceived to be more of an immoral person (in comparison with an agent who attempts and fails to help). This finding replicated even in situations in which it was unlikely that the agent would succeed in helping and was independent of the initial perceived morality of the agent. Moreover, other important cognitive factors such as causal ascriptions, intentions, and justification also explained the omission effect on blame, as predicted by Malle et al. (2014). Hence, this study provided evidence that choosing to omit help receives greater blame in part because individuals are perceived to be bad when they make such decision.

The second examination was conducted to understand whether the inferences of moral character explain the effect of mental states on blame. Importantly, the literature on the positive-negative asymmetry of several psychological processes suggests that the effect of negative mental states should be greater than that of positive mental states. Hence, first, a study was conducted to investigate the hypothesis that intentions predict evaluations of character qualified by a negativity bias. The results revealed that individuals who desired to do good or harm and took action to fulfill their intentions received the greatest ratings for moral and immoral character, respectively. Intentions that were unfulfilled, that is, followed by inaction, due to an external explanation were also more diagnostic of moral character compared to intentions that did not present an external explanation, and this effect was also stronger when the intentions were negative (vs. positive). In other words, individuals that desired to do harm and were about to do it (i.e., suggesting intentionality) were considered to be more immoral than individuals that desired to do good and were about to do it were considered to be moral. This asymmetric effect was explained by the particular informativeness of negative intentions. Negative intentions that are

consistent and distinct from others are more telling of disposition than consistent and distinct positive intentions, likely because positive intentions are taken as the default mental state.

Finally, I examined whether the moral character inferred by intentions (in the study, explored in the form of desires) explains judgments of blame for failing to keep a promise. This study revealed that individuals with negative intentions were deemed as more blameworthy for failing to keep a promise compared to those with positive intentions. On the other hand, those with positive intentions had their blame discounted. Both exacerbation (for negative desires) and mitigation (for positive desires) of blame were explained by inferences of character, even after taking into account other potential explanations (i.e., wrongness, causality and intentionality).

The empirical findings presented in this thesis suggest that blame for failure to help rely at least in part on inferences of moral character. Such inferences may be driven either by the agent's behavior (in which omissions drive greater immoral character inference than commissions) or by the agent's perceived mental states (in which more negative mental states drive greater immoral character inference). This thesis discussed consistently the case of the teenagers who filmed and laughed at a man drowning. The teenagers had no legal obligation to save Jamel Dunn, and therefore they did not receive legal punishment for failing to help the man. However, there was a great deal of criticism from the public and punishment was ensued in indirect ways, to the point of occurrences of death threats against the teenagers. One way through which people may blame these teenagers is by focusing on the fact that they failed to take action to help. It is unlikely that the teenagers would have been criticized had they, for instance, called the police after witnessing the situation. Another way which people may blame is by inferring their mental states. The fact that they filmed and laughed seems suggestive of a lack of intentions to help, which should increase blame. The studies on this thesis suggest that, through either way, a central reason why people blamed these teenagers is because they perceived them to be especially immoral

individuals. As shown in Study 1.3, immoral (vs. moral) characters received more blame for their failure to help. People are motivated to identify the moral character of others, so they can establish who is friend and who is a foe and segregate the potentially dangerous immoral individuals from society (Uhlmann et al., 2014).

There are several contributions of the studies presented in this thesis. First, Study 1 contains an important methodological contribution for studies on omission. Research on the omission bias show that individuals who do not take action to avoid an outcome (e.g., allowing an individual to get killed) are typically less blameworthy than those who take action (e.g., actively killing). However, it is also the case that omitting behavior still elicits a degree of blame, which is unaddressed by these studies. In Study 1, I proposed a novel operationalization of omission (i.e., comparing omitted help with failed attempt to help, instead of commission of negative acts) to capture the particular effect of the choice of omitting help on blame. In other words, the classic omission bias reported in literature typically compares inaction (e.g., letting a man drown by taking no action) and action (e.g., drowning a man) in the context of proscriptive immoralities (i.e., drowning a man, that is, committing a negative act that should be inhibited in society). In this study, I propose an investigation comparing inaction (e.g., letting a man drown by taking no action) and action (e.g., attempting to save the man yet failing) in the context of prescriptive immoralities (i.e., failing to help, in which helping is a behavior that should be promoted in society). This investigation focused on prescriptive immoralities allowed to explore the effect of omission per se.

The findings in this thesis also suggest that judgments of blame for omitted help are likely less structured compared to judgments of commission proposed by stepwise theories of blame (Shaver, 1985; Malle et al., 2014). For example, from Malle et al.'s (2014) perspective, intentional acts lead to search for justification information ("Why did the agent perform this

intentional act?"), whereas unintentional acts should elicit judgments of obligation and capacity ("Was the agent obligated to avoid the outcome and were they capable of avoiding it?"). However, in Study 1, an individual who chose to omit help versus taking action to help was perceived as having less justifications for their failure to help, which explained greater blame for this condition. That is, justification assessment was influenced by the omission manipulation even though the omission was *factually* unintentional. However, people may have inferred both a level of intentionality and unintentionality in the scenarios of failure to help, stemming judgments from both pathways, instead of only one way. As Malle et al.'s (2014) propose, the hierarchical structure of judgments of blame may loosen as the judgments become more complex. Further studies may explore the extent to which these judgments occur in a structured order, such as in Monroe and Malle (2018).

The empirical findings of this thesis also cast light into the underlying processes of blame for situations in which the agent is not causal nor intentional of an outcome. Study 1 showed scenarios in which the agent did not have a causal and intentional role in the outcome (e.g., seeing a man drown and being physically incapable of helping), and revealed that participants still inflated the inferences of causality and intentionality when the agent decided to do nothing. These findings support Alicke's (1992) proposition that negative spontaneous evaluations may motivate individuals to blame, inflating other types of moral judgments, such as causality and intentionality. Study 3 revealed that individuals were more blameworthy for a failure to keep a promise when they possessed negative (vs. positive) desires, and ascriptions of causality and intentionality did not explain this effect of desires on blame. Theories on blame have discussed the centrality of causality and intentionality on blame. This study, just as Inbar et al.'s (2012), show empirical evidence that blame can also be ascribed in the absence of these factors, as long as there is inference of character immorality.

Finally, this study shows support for the negative-positive asymmetry in person perception. In Study 2, negative intentions were more diagnostic of character than positive intentions. In Study 1, taking action to help (vs. omitting it) diminished judgments of immoral character when the agent had already a good and neutral disposition, yet an agent who presented immoral disposition did not have his immorality discounted after the information that he took action to help.

In this thesis, I reviewed important theories on blame. Some theories suggest that the process of assigning blame occurs in a stage-like manner, with careful consideration of several cognitive elements adopted in criminal law, such as Shaver's (1985) TB and Malle et al.'s (2014) PMB. Other theories suggest that the process of blaming is influenced by motivated reasoning, based on heuristics such as assumptions of moral character. Whereas there is ample support for these theories in their own respect, there is still little work testing both the law-based and the character-based aspects of blame at once. In fact, Malle's et al.'s (2014) PMB and Alicke's (2000) CCM are two competing approaches to blame. In this thesis, I aimed to initiate a tentative integration of these distinct theories at the empirical level. The findings presented in this thesis suggest that both aspects may play a concomitant role in influencing blame, at least in the context of failure to help.

Limitations, future directions, and implications

These studies present several limitations. First, all studies used the vignette technique to investigate the blame for failure to help. It is important to complement the present findings with other methods, such as social experiments in which participants are exposed to failure to help and ascribe blame in real time in a "real" setting (e.g., experiments using economic games).

Moreover, the data was collected only with Japanese participants, hence it may not be generalizable to other cultures. For example, it is possible that the role of moral character on

blame found in the present study may be attenuated in cultures which prioritize focus on outcomes, such as those with doctrines of Opacity of Mind (McNamara et al, 2019). The norms of obligation to help also varies across cultures (Baron & Miller, 2000; Butchel et al., 2018), hence future studies may explore the current findings in different cultural settings.

There are several interesting paths to explore which may strengthen the moral character argument in blaming failure to help. Future studies may explore individual differences in such judgements. For instance, individuals with reduced empathy show greater propensity to endorse in utilitarian reasoning (Patil & Silani, 2014). It is possible that the role of moral character in blame for failure to help may be attenuated for such individuals. Research also shows differences in moral judgments depending on one's political beliefs. Liberals tend to rely on concerns regarding harm and fairness, whereas conservatives rely on these and authority, purity and loyalty concerns (Graham et al., 2009). Failing to help an authority, for instance, may designate stronger judgments of blame for conservatives (vs. liberals). Future studies may explore such individual differences and its effects on blame for failure to help.

The present studies did not distinguish between different relationships between the agent and the "victim". This is an interesting venue to pursue, because ascriptions of moral character are likely determined by the nature of the relationship between the parties (Janoff-Bulman & Carnes, 2003; Haidt & Baron, 1996). For example, people judged an ambiguous behavior to be less likely to constitute a transgression when the target was one's brother (vs. stranger), revealing that moral reasoning depends on the relationship between parties. In the context of failure to help, it seems reasonable that failing to help one's family member stems more blame than failing to help a stranger. This judgment may stem from perceptions of obligation, but it may also be the case that people may blame on the perception of particular immoral character ("only a terrible person would not help their own brother"). Likewise, a similar pertinent path of investigation is

to explore the morality of the victim. In Study 1.3, we manipulated the moral disposition of the agent who failed to help, and the results showed that more immoral agents received greater blame. It is likely that the morality of the victim also affects judgments of blame (e.g., they likely distinguish blame for failing to help a moral man versus a criminal). Again, who the "victim" is should matter little to rational judgments of blame, yet a person-based approach predicts a motivation to blame immoral characters. It is possible that people may increase justifications for failure to help immoral characters. Another option is that such effect could be explained, or perhaps interact, with group identification. As shown in Monroe and Malle (2018), blame judgments were harsher when directed at members of an outgroup. Future studies may examine these interesting possibilities, for example, by examining how people assign blame for failure to help in the context of groups and institutions.

In terms of the psychological processes, it is also relevant to observe whether the character explanation persists under warrant. In the studies of this thesis, participants blamed agents without any consequence to their judgments. In real life, however, blame can be costly, and unfair blaming is backlashed (Malle et al., 2014). Future studies may explore whether the processes underlying blame for failure to help changes when individuals are held accountable for their judgment.

The studies in this thesis show that failing to take action to help or displaying negative mental states are both forms of diagnosing immoral character, which in turn invite judgments of blame. There are important implications in the real-world regarding these findings. From the legal perspective, individuals should not be punished for who they are, but for their criminal behavior. However, that's not how people seem to blame others. The case of the teenagers from Dunn's case is a seeming example of how people blame failure to help motivated to punish individuals, rather than their behaviors.

The motivation to help others comes not only from altruism, moral duty, or to avoid negative outcomes, but also stems from impression management. Failing to help may cost one's reputation. For example, in the recent events of manifestations for the Black Lives Matter in 2020, there were movements to boycott companies that did not support the movement. Exploring the psychological underpinnings of blame for omission in corporations is an important investigation with real-world implications to areas of business management, marketing and entrepreneurship.

Daily life interactions show a great deal of examples of blaming failure to help through underlying perceptions of character. Parents blame their children for being unhelpful by stating they are "bad", people easily assume that one who fails to help another in an emergency must be "psychopathic", partners blame each other for failing to support one another by assuming the partner is "selfish". Movies highlight the moral character of the hero who attempts to save the victim even when he knows he won't make it. Likewise, the hero who omits help is frequently portrayed with guilt and shame for his own lack of moral character. With his theory of lame, Shaver (1985) aimed to provide a framework for people to evaluate their own blame processes. Understanding that people blame failure to help based on assumptions of character can further help individuals to think critically about their own processes of blame. For example, media typically portraits individuals engaging in heroic acts to help others (even when the chances of succeeding are very unlikely), however individuals in real life are very likely to omit help, as shown, for instance, in the bystander effect (Latané & Darley, 1968). It is then relevant to question to which extent explaining blame through attributions of character is sensible. Suppose that, in a clinical setting, a mother who failed to help her child comes with intense feelings of shame and guilt because she thinks of herself as a horrible mother. Reanalyzing the process

considering other factors (such as whether she was capable of helping or had the obligation to help) is important to reframe the narrative of blame.

Summary

The studies on morality have long focused on the processes through which people blame proscriptive immoralities, that is, the failure to inhibit harmful behavior. The present thesis investigated the psychological processes underlying people's judgments of blame for prescriptive immorality, that is, when an agent fails to help. Two important information are used to form such judgments: the agent's actions or inactions and the agent's mental states. The empirical work of this thesis reveals that an agent's actions and mental states influence blame in part because both are indicative of their underlying moral character. Stepwise blame models do not consider the role of moral character on blame (e.g., Shaver, 1985: Theory of Blame) or tend to refute it (Malle et al., 2014) in favor of other cognitive aspects, such as causality and intentionality. The empirical findings of this thesis revealed that the moral character explanation persisted after taking into account other potential explanations, such as causality and intentionality. Moral character seems to be an important explanation to the well-known effect of both action and mental states on blame, at least in the context of failure to help, such as omissions.

References

Albarracín, D., Sunderrajan, A., Dai, W., & White, B. X. (2019). The social creation of action and inaction: From concepts to goals to behaviors. In J. M. Olson (Ed.), *Advances in experimental social psychology: Vol. 60. Advances in experimental social psychology* (p. 223–271). Elsevier Academic Press.

Algoe, S. B., & Haidt, J. (2009). Witnessing excellence in action: The "other-praising" emotions of elevation, gratitude, and admiration. *The Journal of Positive Psychology, 4*(2), 105–127.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*(3), 368-378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126,* 556–574. https://doi.org/10.1037//0033-2909.126.4.556

Alicke, M. D. (2008). Blaming badly. *Journal of Cognition and Culture, 8,* 179–186.

Alicke, M. D., & Davis, T. L. (1989a). The role of a posteriori victim information in judgments of blame and sanction. *Journal of Experimental Social Psychology, 25,* 362–377.

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *Perspectives on Psychological Science, 10*(6), 790–812. https://doi.org/10.1177/1745691615601888

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *Journal of Philosophy,* 108, 670–696.

Alicke, M. D., & Zell, E. (2009). Social attractiveness and blame. *Journal of Applied Social Psychology, 39,* 2089–2105.

Ames, D., & Fiske, S. T. (2013). Intentional harms are worse, even when they're not. *Psychological Science, 24*(9), 1755-1762.

Apperly, I. (2011). *Mindreaders: The cognitive basis of the "Theory of Mind".* Psychology Press.

Avramova, Y., & Inbar, Y. (2013). Emotion and moral judgment. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(2), 169-178.

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

Baron, J., & Miller, J. G. (2000). Limiting the scope of moral obligations to help: A cross-cultural investigation. *Journal of Cross-Cultural Psychology, 31*(6), 703–725. https://doi.org/10.1177/0022022100031006003

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes, 94*(2), 74–85. https://doi.org/10.1016/j.obhdp.2004.03.003

Baron, J., & Ritov, I. (2009). Chapter 4 Protected Values and Omission Bias as Deontological Judgments. *Psychology of Learning and Motivation - Advances in Research and Theory*, Vol. 50, pp. 133–167. https://doi.org/10.1016/S0079-7421(08)00404-0

Barrett, L, Dunbar, R, & Lycett, J. (2002). *Human Evolutionary Psychology*. Princeton University Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. & Vohs, K. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323-370.

Bocian, K., Baryla, W., Kulesza, W. M., & Schnall, S. (2018). The mere liking effect: Attitudinal influences on attributions of moral character. *Journal of Experimental Social Psychology, 79*, 9-20.

Bogardus, E. S. (1933). A social distance scale. *Sociology & Social Research*, 17, 265–271.

Boster, F. J., Fediuk, T. A, & Kotowski, R. (2001). The effectiveness of an altruistic appeal in the presence and absence of favors. *Communication Monographs, 68*(4), 340–346.

Bostyn, D. H., & Roets, A. (2016). The morality of action: The asymmetry between judgments of praise and blame in the action-omission effect. *Journal of Experimental Social Psychology, 63*, 19–25. https://doi.org/10.1016/j.jesp.2015.11.005

Branscombe, N.R., Owen, S., Garstka, T.A. & Coleman, J. (1996). Rape and accident counterfactuals: Who might have done otherwise and would it have changed the outcome?. *Journal of Applied Social Psychology, 26,* 1042–67.

Buchtel, E. E., Ng, L. C. Y., Norenzayan, A., Heine, S. J., Biesanz, J. C., Chen, S. X., Bond, M. H., Peng, Q., & Su, Y. (2018). A sense of obligation: Cultural differences in the experience of obligation. *Personality and Social Psychology Bulletin, 44*(11), 1545–1566. https://doi.org/10.1177/0146167218769610

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLoS One, 10*(8), 1–20. https://doi.org/10.1371/journal.pone.0136589

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2), 284-299.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition, 136,* 30-37.

Cheng, P. W., & Novick, L.R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology,* 58(4), 545-567.

Choi, I., Nisbett, R.E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin, 125*(1), 47-63.

Christopher, R. (2004). Does attempted murder deserve greater punishment than murder – moral luck and the duty to prevent harm. *Notre Dame Journal of Law, Ethics & Public Policy, 18*(2), 419-435.

Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology, 81*(4), 697-710.

Constanzo, P. R., Coie, J. D., Grumet, J. F., & Farnill, D. (1973). A reexamination of the effects of intent and consequence on children's moral judgments. *Child Development, 44*(1), 154-161.

Bode, N. W. F., Miller, J., O'Gorman, R. & Codling, E. A. (2015). Increased costs reduce reciprocal helping behaviour of humans in a virtual evacuation experiment. *Scientific Reports, 5,* 15896..

Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality,* Vol. 78, pp. 106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353-380.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*(6), 1052–1075. https://doi.org/10.1111/j.1551-6709.2010.01167

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089.

DeScioli, P., Asao, K., & Kurzban, R. (2012). Omissions and Byproducts across Moral Domains. *PLoS ONE, 7*(10). https://doi.org/10.1371/journal.pone.0046963

DeScioli, P., Bruening, R., & Kurzban, R. (2011). The omission effect in moral cognition: toward a functional explanation. *Evolution and Human Behavior, 32*(3), 204-215. https://doi.org/10.1016/j.evolhumbehav.2011.01.003

Dovidio, J. F., Piliavin, J. A., Schroeder, D. A., & Penner, L. A. (2006). *The social psychology of prosocial behavior*. Erlbaum.

Dressler, J. (2015). *Understanding criminal law* (7th ed.). LexisNexis.

Fazio, R.H., Sanbonmatsu, D.M., Powell, M.C., & Kardes, F.R. (1986). On the automatic activation of attitudes. *Journal of Personality and Social Psychology, 50*, 229-238.

Fincham, F. D., & Jaspars, J. M. (1979). Attribution of responsi- bility to the self and other in children and adults. *Journal of Personality and Social Psychology, 37*, 1589–1602.

Fincham, F. D., & Jaspars, J. M. (1980). Attribution of responsibility: From man the scientist to man as lawyer. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 13, pp. 81–138). Academic Press.

Forgas, J.P., & Bower, G. (1987). Mood effects on person-perception judgments. *Journal of Personality and Social Psychology, 53*(1), 53-60.

Forsterling, F. (1989). Models of Covariation and attribution: How do they relate to the analogy of Analysis of Variance?. *Journal of Personality and Social Psychology*, 57(4), 615-625.

Gallop, J. D. (2018, July). *Teens mock drowning man while filming him: Cocoa tragedy still resonates with families 1 year later.* Florida Today.

Goldberg, J. H., Lerner, J. S., & Tetlock, P. E. (1999). Rage and reason: the psychology of the intuitive prosecutor. *European Journal of Social Psychology, 29*(56), 781–795.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). *Chapter Two - Moral foundations theory: The pragmatic validity of moral pluralism Advances in Experimental Social Psychology*, 47, 55-130. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101-124.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Gromet, D. M., Goodwin, G. P., & Goodman, R. A. (2016). Pleasure from another's pain: The influence of a target's hedonic states on attributions of immorality and evil. *Personality and Social Psychology Bulletin, 42*(8), 1077–1091.

Guglielmo, S. (2012). *The information-seeking process of moral judgment.* Unpublished dissertation. Brown University, Providence, RI.

Guglielmo, S., & Malle, B. F. (2014). *Information-seeking processes in moral judgment.* Manuscript in preparation.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814-834.

Haidt, J. (2003). The moral emotions. In R.J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 852-870). Oxford: Oxford University Press.

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316,* 998-1002.

Haidt, J. & Baron, J. (1996). Social roles and the moral judgement of acts and omissions. *European Journal of Social Psychology,* 26(2), 201-218.

Haidt, J., & Joseph, C. (2008). The moral mind: How five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers,

S. Laurence, & S. Stich (Eds.), *Evolution and cognition. The innate mind Vol. 3. Foundations and the future* (p. 367–391). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195332834.003.0019

Hamilton, V.L. (1980). Intuitive psychologist or intuitive lawyer? Alternative models of the attribution process. *Journal of Personality and Social Psychology, 39*(5), 767-772.

Hamlin, J. K. (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition*, 128, 451–474.

Hart, H. L. A., & Honoré, A. M. (1959). *Causation in the law*. Oxford, England: Oxford University Press.

Hayashi, H. (2015). Omission bias and perceived intention in children and adults. *British Journal of Developmental Psychology,* 33, 237-251. https://doi.org/10.1111/bjdp.12082

Hayes, A. F. (2013). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach.* The Guilford Press.

Heider, F. (1958). *The psychology of interpersonal relations.* Wiley.

Hirozawa, P. Y., Karasawa, M., & Matsuo, A. (2019). Intention matters to make you (im)moral: Positive-negative asymmetry in moral character evaluations. *Journal of Social Psychology,* 1–15. https://doi.org/10.1080/00224545.2019.1653254

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin, 38*(1), 52-62.

Janoff-Bulman, R., & Carnes, N. C. (2013). Surveying the Moral Landscape: Moral Motives and Group-Based Moralities. *Personality and Social Psychology Review, 17*(3), 219–236. https://doi.org/10.1177/1088868313480274

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, *96*(3), 521–537. https://doi.org/10.1037/a0013779

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in social psychology. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 219-266). New York: Academic Press.

Jones, E. E., & Nisbett, R. E. (1972). The actor and the observer: Divergent perceptions of the causes of behavior. In E. E. Jones, D. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 79 –94). General Learning Press.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review, 93*, 136–153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge University Press.

Karlovac, M., & Darley, J. M. (1988). Attribution of responsibility for accidents: A negligence law analogy. *Social Cognition*, *6*(4), 287–318. https://doi.org/10.1521/soco.1988.6.4.287

Keiler, J., Panzavolta, M., & Roef, D. (2017). *Criminal law*. In J. Hage, A. Waltermann, & B. Akkermans (Eds.), *Introduction to Law* (2nd ed., pp. 129-164). Springer. https://doi.org/10.1007/978-3-319-57252-9

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska Symposium on Motivation* (pp. 192-238). Lincoln: University of Nebraska Press.

Kensinger, E. A., Garoff-Eaton, R. J., & Schacter, D. L. (2006). Memory for specific visual details can be enhanced by negative arousing content. *Journal of Memory and Languag*e, *54*, 99-112.

Klein, J. (1991). Negativity effects in impression formation: A test in the political arena. *Personality and Social Psychology Bulletin, 17*(4), 412-418.

Kleinke, C. L., Wallis, R., & Stadler, K. (1992). Evaluation of a rapist as a function of expressed intent and remorse. *The Journal of Social Psychology, 132*, 525-537

Kordes-de-Vaal, J. (1996). Intention and omission bias: Omissions perceived as nondecisions. *Acta Psychologica, 93*(1-3), 161-172.

Lagnado, D., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*(3), 754-770. https://doi.org/10.1016/j.cognition.2008.06.009

Latané, B., & Darley, J. M. (1968). Group inhibition of bystander intervention in emergencies. *Journal of Personality and Social Psychology, 10*(3), 215-221.

Malle, B. F. & Bennett, R. (2002). People's praise and blame for intentions and actions: Implications of the folk concept of intentionality. *Annual meeting of the American Psychological Association.*

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*(2), 101–121. https://doi.org/10.1006/jesp.1996.1314

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-Observer Asymmetries in Explanations of Behavior: New Answers to an Old Question. *Journal of Personality and Social Psychology, 93*(4), 491–514. https://doi.org/10.1037/0022-3514.93.4.491

Mazzocco, P. J., Alicke, M. D., & Davis, T. L. (2004). On the Robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology*, *26*(2–3), 131–146. https://doi.org/10.1207/s15324834basp2602&3_3

McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. *Cognition*, *182*, 95–108. https://doi.org/10.1016/j.cognition.2018.09.008

Mende-Siedlecki, P., Baron, S., & Todorov, A. (2013). Diagnostic value underlies asymmetric updating of impressions in the morality and ability domains. *Journal of Neuroscience, 33*(50), 19406-19415.

Mill, J. S. (1998). *Utilitarianism* (R. Crisp, Ed.). Oxford University Press.

Monroe, A. E., & Malle, B. F. (2018). Two paths to blame: intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General*, *146*(1), 123-133.

Montoya, A., & Hayes, A. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*(1), 6-27.

Moussaïd, M., & Trauernicht, M. (2016). Patterns of cooperation during collective emergencies in the help-or-escape social dilemma. *Scientific Reports*, 6, 33417. https://doi.org/10.1038/srep33417

Mullen, E., & Skitka, L. J. (2006). Exploring the psychological underpinnings of the moral mandate effect: motivated reasoning, group differentiation, or anger? *Journal of Personality and Social Psychology*, 90, 629-643.

Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law and Contemporary Problems, 75*(2), 1 – 31.

Nadler, J. (2014). The path of motivated blame and the complexities of intent. *Psychological Inquiry, 25*(2), 222-229.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology, 80,* 381–396.

Patil, I., & Silani, G. (2014). Reduced empathic concern leads to utilitarian moral judgments in trait alexithymia. *Frontiers in Psychology, 5 (*501), 1-12.

Peeters, G., & Czapinsky, J. (1990). Positive-negative asymmetry in evaluations: the distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*(1), 33-60.

Piaget, J. (1965). *The moral judgment of the child*. New York: Free Press.

Piazza, J., Souza, P., Rottman, J., & Syropoulos, S. (2018). Which appraisals are foundational to moral judgments? Harm, injustice and beyond. *Social Psychological and Personality Science*, ISSN 1948-5506.

Pizarro, D., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In P. Shaver & M. Mikulincer (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). APA Books.

Pizarro, D. A., Tannenbaum, D., & Uhlmann, E. (2012). Mindless, harmless, and blameworthy. *Psychological Inquiry, 23*(2), 185–188. https://doi.org/10.1080/1047840X.2012.670100

Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14*(3), 267-272.

Premack, D. & Wooduff, G. (1978). Does the chimpanze have a theory of mind? *The Behavioral and Brain Sciences, 4,* 515-526.

Rand, D. G., & Epstein, Z. G. (2014). Risking your life without a second thought: Intuitive decision-making and extreme altruism. *Public Library of Science One, 9*(10): e109687, 1-6.

Rand, D., Peysakhovich, A., Kraft-Todd, G., Newman, G., Wurzbacher, O., Nowak, M. & Greene, J. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*(1), 1-12.

Reeder, G., & Brewer, M. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86(*1), 61-79.

Reeder, G., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology, 44*(4), 736-745.

Riskey, D. R., & Birnbaum, M. H. (1974). Compensatory effects in moral judgment: Two rights don't make up for a wrong. *Journal of Experimental Psychology, 103*(1), 171-173.

Ritov, I., & Baron, J. (1990). Reluctance to vaccinate: Omission bias and ambiguity. *Journal of Behavioral Decision Making, 3*(4), 263–277. https://doi.org/10.1002/bdm.3960030404

Ritov, I., & Baron, J. (1999). Protected values and omission bias. *Organizational Behavior and Human Decision Processes, 79*, 79–94.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.

Salerno, J., & Peter-Hagene, L. (2013). The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science, 24*(10), 2069-2078.

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty, 1,* 7–59.

Schleifer, M., Shultz, T. R., & Lefebvre-Pinard, M. (1983). Children's judgements of causality, responsibility and punishment in cases of harm due to omission. *British Journal of Developmental Psychology, 1*(1), 87–97.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness.* Springer Verlag.

Shultz, T. R., Schleifer, M., & Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Canadian Journal of Behavioural Science, 13*(3), 238–253. https://doi.org/10.1037/h0081183

Shultz, T. R., & Wright, K. (1985). Concepts of negligence and intention in the assignment of moral responsibility. *Canadian Journal of Behavioural Science,* 17(2), 97–108.

Shweder, R. A., Much, N. C., Mahapatra, M., & Park, L. (1997). The "big three" of morality (autonomy, community, divinity) and the "big three" explanations of suffering. In A. M. Brandt & P. Rozin (Eds.), *Morality and health* (p. 119–169). Taylor & Frances/Routledge.

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

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*(1), pp. 76–105. https://doi.org/10.1016/0022-1031(91)90011-T

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, 47, 1249–1254.

Ugazio, G., Lamm, C., & Singer, T. (2012). The role of emotions for moral judgments depends on the type of emotion and moral scenario. *Emotion, 12*(3), 579-590.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81. https://doi.org/10.1177/1745691614556679

Uhlmann, E. L., Zhu, L., & Diermeier, D. (2014). When actions speak volumes: The role of inferences about moral character in outrage over racial bigotry. *European Journal of Social Psychology*, 44, 23–29.

Uhlmann, E. L., Zhu, L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, 126, 326–334.

Valencia, N. & Sayers, D. (2017, July). *Florida teens who recorded drowning man will not be charged in his death.* CNN.

Van de Ven, N., Archer, A. T. M., & Engelen, B. (2018). More important and surprising actions of a moral exemplar trigger stronger admiration and inspiration, *The Journal of Social Psychology*, DOI: 10.1080/00224545.2018.1498317

Wells, G.L. & Gavanski, I. (1989). Mental simulation of causality. *Journal of Personality and Social Psychology, 56*,161–9.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition, 120*(2), 202-214.

Appendix 1.1

Study 1.1 Scenarios

Set 1: Failure to help with direct harm to an individual

1. The Swimmer Scenario

Tanaka is walking leisurely at a deserted beach, when he notices a swimmer far away in the sea. The swimmer was clearly drowning. Tanaka notices it was possible for Tanaka to reach the swimmer in time.

Action. Hence, Tanaka chooses to throw himself in the water. He swims, but to no avail. He returns to the shore empty handed.

Inaction. Still, Tanaka chooses to stay, as he thought it was unlikely that he could help. With no rescue, the swimmer eventually dies.

2. The Child Scenario

Sato is walking his dog at a deserted park. He suddenly hears a scream. From afar, he sees a boy hanging from a very tall tree. It was possible that Sato could reach the boy before he would fall.

Action. Hence, Sato chooses to let go of the dog and dashes towards the boy with arms ready to catch him. Nevertheless, he does not make it in time.

Inaction. Still, Sato chooses to just stand there, watching, as he thought there was little he could do.

The boy finally cannot stand it anymore. He falls from the tall tree and suffers a severe head injury, which would affect him for the rest of his life.

3. The Driver Scenario

Matsuo is walking back home in a deserted street, when he hears a faint sound of crying. Matsuo searches and sees a drunk man struggling to put his crying child in the back seat. Matsuo realizes that was a dangerous situation. They weren't far. Matsuo thought it was possible for him to reach and stop them in time.

Action. Hence, Matsuo chooses to run after them to stop the man from driving. However, Matsuo does not reach them in time.

Inaction. Still, Matsuo chooses to stay, as he didn't think that was his business.

The man finally succeeds in putting his child in the car. Soon enough, he starts the car in explosive speed. One kilometer later, the car hits violently against a pole. Both man and child were severely injured.

Set 2: Failure to help with indirect harm through material damage

4. The Fire Scenario

Nakamura was out jogging, when he notices a smoke coming from a nearby, isolated house. Alarmed, he checks from the window. The fire, caused by a candle, was spreading dangerously. The street was deserted, and he had no phone with him. In 15 minutes, the fire would likely consume the house.
Nakamura remembered the closest convenience store was 500 meters away.

Action. Hence, Nakamura chooses to run to the convenience store to get help, but when he gets there it is too late.

Inaction. Still, Nakamura chooses to just stand there, watching, as he thought there was nothing he could do.

Eventually, the whole house was on fire.

5. The Painting Scenario

Itou stops by a local shop from an old artist. The old man shows Itou around the shop and gets to like Itou. Hence, old man decides to show him his secret masterpiece - a gorgeous painting that took him 15 years to finish. Amazed, Itou examines it as the old man leaves to get some tea. Suddenly, the store starts shaking. Itou runs and hides under a sturdy shelf. The earthquake was weak and lasts only a few seconds. However, it was enough to set a large can of black ink over the table to its edge. Itou notices the can and realizes that it was about to fall. If it did, the can would certainly splash and destroy the old man's painting.
The can was close. It was likely that Itou could reach it in time.

Action. Itou chooses to storm towards the can in an attempt to stop it from falling. He did not manage, though.

Inaction. Still, Itou chooses to just stand there, thinking there was nothing he could do.

Eventually, the can falls from the table and black ink splashes everywhere. The masterpiece is ruined.

6. The Computer Scenario

It is raining heavily outside, and Yamazaki is working late at the laboratory. Suddenly, he hears a sound from above. A crack had appeared on the ceiling, from which water would certainly start to pour. Yamazaki notices the computer underneath it - his coworker's computer, which held the data of all of his coworker's hard work from the past two years. Yamazaki's desk was close to his colleague's. It was likely that he could reach it in time.

Action. Hence, Yamazaki chooses to run to the computer in an attempt to save it from the water, but it was too late.

Inaction. Still, Yamazaki chooses to just stand there, watching, thinking there was nothing he could do.

Eventually, an increasing amount of water splashed over the machine, certainly damaging it. His coworker's data was lost.

Appendix 1.2

Study 1.2

The Swimmer Scenario

Nakamura had been practicing swimming at a small lake a few kilometers from his house. His goal was to swim all the way to the other side of the lake. It was a challenging goal, as he would always get fatigued half-way through and had to use his lifejacket to help him return safely. On his 30th training day, Nakamura reached his best performance, which was ¾ of the route.
The next day, Nakamura went again swimming. He entered the water, when he noticed a man clearly drowning at the other side of the lake. There was no time to call for help.
Nakamura wanted to save him. However, he knew he had never been able to reach the other side of the lake before.

> **Action.** Still, Nakamura decided to swim towards the man. However, as expected, he got too fatigued after ¾ of the route and could no longer swim.
>
> **Inaction**. Hence, Nakamura decided not to swim towards the man.

In a couple of minutes, as the man got no rescue, the man drowned to death.

The Child Scenario

Every day, Nakamura had been sprinting. His goal was to run 20m in 5 seconds, yet he would always take longer than that. After two months practice, he reached his best performance, which was 20m in 6 seconds.
The next day, Nakamura went again for his practice. He had just finished warming up, when he heard a child screaming. He noticed it was a child hanging by a very tall tree 20m away. The child was holding a long branch with both arms but was slipping and looked just about to fall.
Nakamura wanted to help. However, he knew he had never been able to sprint that fast before.

> **Action**. Still, Nakamura decided to sprint towards the boy, but to no avail. As expected, he got too fatigued midway and lost his speed.
>
> **Inaction**. Hence, Nakamura decided not to sprint towards the boy.

In a matter of 5 seconds, the boy fell from the tree. The injury left the boy with a lifelong sequelae.

Appendix 1.3

Study 1.3 Moral disposition manipulation

Moral. Nakamura was a kind-hearted man. Since childhood, Nakamura worried about other people's feelings, so he tried his best to listen to others and would not gossip or talk badly about anyone. At work, everyone knew Nakamura was a trustworthy man. Nakamura had never missed a day of work and worked diligently to meet the deadlines he had promised. As he was a good listener, Nakamura was also frequently sought by his coworkers, who trusted him with both work and personal problems.
In his free time, Nakamura spent most of his time with his wife. He had been faithfully married for ten years. Nakamura also enjoyed taking walks with their dog, and from time to time joining a volunteer work from the neighborhood. Recently, Nakamura had also been interested in exercising.

Neutral. Nakamura was an ordinary man. He worked as a researcher and spent his days in his office, reading papers and running experiments. His coworkers thought Nakamura was a quiet man and did not have a particular impression on him. In his free time, Nakamura enjoyed eating at good restaurants, reading a book and sleeping. Recently, Nakamura had also been interested in exercising.

Immoral. Nakamura was an uncaring man. Since childhood, Nakamura was indifferent about other people's feelings, so he frequently ignored others and gossiped and talked badly of them. At work, everyone knew Nakamura was an untrustworthy man. He faked sickness to miss work and would frequently miss deadlines he had promised to meet. His coworkers also avoided talking to Nakamura outside of work. That was because once a coworker told Nakamura that he was in financial difficulties, and the next day Nakamura told everyone about it.
In his free time, Nakamura spent most of his time with his mistress. He had been cheating his wife for ten years. Nakamura also enjoyed overdrinking at downtown pubs while gambling. Recently, Nakamura had also been interested in exercising.

Scenarios

The Swimmer Scenario

From time to time, Nakamura would practice swimming at a small lake a few kilometers from his house. His goal was to swim all the way to the other side of the lake without taking a break to rest. However, as he could not practice consistently, his performance was also inconsistent.
Out of 20 times Nakamura had practiced, 10 times he had been able to cross the lake without taking a break to rest.
One day, Nakamura went again swimming. He entered the water, when he noticed a man clearly drowning at the other side of the lake. There was no time to call for help.
Nakamura thought he only be able to save the man if he could swim all the way without taking a break. However, Nakamura did not know whether he would be able to do so in time.

Action. Still, in the end, Nakamura decided to swim towards the man. However, he got too fatigued and could not reach the man in time.
Inaction. Hence, in the end, Nakamura decided not to swim towards the man.

As the man got no rescue, the man drowned to death.

The Child Scenario

From time to time, Nakamura would practice sprinting. His goal was to run 50m in 8 seconds. However, as he could not practice consistently, his performance was also inconsistent.
Out of 20 times Nakamura had practiced, he had been able reach his goal 10 times.
One day, Nakamura went again training for sprinting. He had just finished warming up, when he heard a child screaming. He noticed a boy was hanging by a very tall tree 20m away. The boy was holding a long branch with both arms but was slipping and looked just about to fall. Nakamura thought he would only be able to save the boy if he would be able to run in his fastest speed record. However, Nakamura did not know whether he would be able to do so.

- **Action**. Still, in the end, Nakamura decided to sprint towards the boy. However, he was not fast enough to reach the boy in time.
- **Inaction**. Hence, in the end, Nakamura decided not to sprint towards the boy.

In a matter of 8 seconds, the boy fell from the tree. The injury left the boy with a lifelong sequelae.

Appendix 2

Description of Positive Vignettes from Study 2.2

Positive vignettes				
Vignettes	Description	Unfulfilled intention unexplained	Unfulfilled intention explained	Fulfilled intention
1. Donating money	Tanaka was walking across the university hall, when (s)he saw a donation campaign flyer glued to the wall and a donation box. It was about a university organization which was collecting money to donate to underprivileged children. Tanaka thought it was very important to help people in need...	...so (s)he took his/her wallet out while thinking of donating. Nevertheless, although (s)he was still thinking of donating, Tanaka ended up passing by the box.	...so (s)he took his/her wallet out while thinking of donating. Nevertheless, when (s)he opened his/her wallet, (s)he found it empty. Tanaka had forgotten that, right before going to the university, (s)he had spent all the money with groceries. Therefore, Tanaka passed by the box without donating.	...so (s)he wanted to donate. Therefore, (s)he took his/her wallet and placed the money inside the donation box.
2. Helping a classmate with math	Nakamura heard that a colleague from the same class was having difficulties with math. Nakamura was good with math and wanted to somehow support that classmate...	...Therefore, Nakamura intended to use as much as possible of his/her free time during the weekend to teach the classmate, even changing his/her initial plans for the weekend. Nevertheless, the weekend passed without Nakamura meeting his/her classmate.	...Therefore, Nakamura planned to use as much as possible of his/her free time during the weekend to teach the classmate, even changing his/her initial plans for the weekend. Nevertheless, a warning of typhoon was announced on that weekend, making it unsafe to leave the house. Thus, the weekend passed without Nakamura meeting the classmate.	...Therefore, (s)he used as much as possible of his/her free time during the weekend to teach his/her classmate.
3. Offer to carry weight	Yamada was crossing the street when (s)he saw an old man carrying some heavy groceries bags. The old man looked like he was having trouble. Yamada thought it was very important to help people in need...	...so (s)he walked to the old man while thinking of offering assistance. Nevertheless, even though still thinking of offering help, Yamada ended up passing by the old man.	...so (s)he walked to the old man while thinking of offering assistance. When Yamada was about to talk to the old man, a friend at the other side of the street called out, and Yamada ended up not talking with the old man.	...so (s)he decided to offer assistance. Therefore, Yamada walked to the old man and offered to help carrying the bags.

Description of Negative Vignettes from Study 2.2

		Negative vignettes		
Vignettes	Description	Unfulfilled intention unexplained	Unfulfilled intention explained	Fulfilled intention
1. Stealing money	Nakamura was around campus, when (s)he saw a forgotten wallet over a table from the common area. Nakamura checked the wallet and noticed there was money inside. Feeling lucky, Nakamura looked around, confirming (s)he was not seen, while thinking of stealing the money from the wallet...	...Nevertheless, in the end Nakamura ended up leaving the study hall without taking any money.	...By the time Nakamura was about to take the money, a friend that had just entered the room called out to him/her. Nakamura ended up not taking any money.	...Furtively, Nakamura took the money and left the wallet over the table.
2. Damaging a classmate's study material	When Takahashi went to the toilet at the university, (s)he found some notes forgotten over the sink. Takahashi took a look at it and realized they were lecture notes from a classmate who Takahashi strongly disliked. Those notes also contained important content for the upcoming test. Takahashi thought of troubling the classmate...	...so (s)he decided to throw those notes away. Nevertheless, in the end Takahashi went out of the toilet leaving the notes where they were.	...so (s)he decided to throw those notes away. Nevertheless, when (s)he was about to do it, some people entered the toilet. Takahashi took his/her hands off the notes and ended up exiting the toilet, leaving the notes where they were.	...so (s)he ended up throwing the notes away.
3. Imposing weight	Matsumoto worked at a supermarket. Part of his/her job consisted of working in cooperation with another staff in organizing the delivered goods. (S)He was separating boxes into two piles, which (s)he and another old man would have to carry one by one to the storage room. Matsumoto disliked the old man, so (s)he decided to place all heavier boxes in the co-worker's pile.	...Nevertheless, Matsumoto ended up not putting the heavier boxes in the co-worker's pile.	…Nevertheless, when (s)he was about to do as intended, the store manager entered the workplace and helped him/her with the sorting process, and thus the heavier boxes were not placed in the co-worker's pile.	...Then, Matsumoto did as intended, making the old man carry all the heavier boxes while getting the lighter ones to himself/herself.

Appendix 3

Study 3. Scenarios

1. Ito's best friend was running a donation project for the homeless. The project was very precious to Ito's friend, and Ito's friend would be very grateful if Ito could contribute.

Positive desires. Ito really wanted to donate. (S)he greatly empathized with homeless people and wished the project would be successful and useful to many.

Neutral desires. Ito didn't particularly want to donate but didn't mind donating either. (S)he didn't like nor dislike the project and was overall neutral towards it.

Negative desires. Ito really did not want to donate. S(h)e felt disgusted by homeless people and wished the project would not be successful so that homeless people would learn to make a living on their own.

On the day of the donation, Ito arrived late, just as the donation stall was about to close. Ito ran to the stall and quickly took his/her wallet to donate. However, as (s)he opened it, Ito realized (s)he had spent all the money on groceries. Hence, despite Ito's promise, (s)he ended up not donating. Days later, Ito found out Ito's best friend had to cancel the project due to lack of donations. "If only I had gotten one more donation, the project would have kept going", the friend told Ito with a sad voice.

2. Nakamura and his/her sister were planning their friend's birthday celebration. The sister wanted to do something special, as their friend had been very sad lately. Then, the sister had the idea of surprising their friend with a very fancy birthday cake after dinner. They knew their friend had always wanted to try that cake but would not buy it as it was too expensive. As the sister would cook the dinner, Nakamura's sister asked Nakamura to buy the cake, and they would split the expenses.

Positive desires. Nakamura really wanted to buy the cake. Nakamura loved celebrations and wished the friend would be happy with the surprise.

Neutral desires. Nakamura didn't particularly want to buy the cake but didn't mind buying it either. (S)he didn't like nor dislike celebrations and was overall neutral about it.

Negative desires. Nakamura really did not want to buy the cake. (S)he hated celebrations and wished the friend would overcome their emotions by themselves.

[However], According to the plan, Nakamura promised to buy the cake.

On the birthday night, Nakamura went to the cake shop after work, but it was closed. (S)he had made a mistake; Nakamura thought the store would be open until 7pm, but it closed at 6pm.

In the end, as they had no surprise cake, the dinner went by flat and did not feel like a celebration at all. Both friend and sister were very disappointed but didn't say anything.

3. Sato's cousin was moving out from their apartment, but the cousin didn't have a car. Sato had a very big car, so his/her mother told him/her to help the cousin.

Positive desires. Sato really wanted to help the cousin. (S)he was happy to be useful for others and wished the cousin would have a smooth moving out process.

Neutral desires. Sato didn't particularly want to help the cousin but didn't mind it either. Sato didn't like nor dislike cooperating with others and had no special opinion about it.

Negative desires. Sato really did not want to help the cousin. (S)he was annoyed to be bothered by others and for that Sato wished the cousin would have a stressful moving out process.

In the end, in line with his/her mother's request, Sato promised the cousin to help with the moving process. The cousin showed great gratitude.

On the moving out day, at the settled time, Sato was about to go to the cousin's apartment, when Sato remembered there was an important meeting at work. Sato rushed to work and ended up cancelling on the cousin.

At night, as Sato was returning home, Sato received another call. His/her mother said the cousin was not able to finish moving out by himself/herself. Hence, the cousin had to pay an expensive fee to the real state agency, as there was still some furniture in the apartment on the day (s)he should have returned the apartment.

4. Yamada's classmate was having lots of difficulty with math. Yamada was very good at math, so Yamada's teacher said it would be nice if Yamada could support the classmate.

Positive desires. Yamada really wanted to help the classmate. (S)he thought that people should always help each other and wished the classmate would persevere on his/her studies.

Neutral desires. Yamada didn't particularly want to help the classmate but didn't mind helping either. (S)he didn't like nor dislike cooperating with others and had no special opinion about it.

Negative desires. Yamada really did not want to help the classmate. (S)he thought people should not be inconvenient to others and wished the classmate would just give up on his/her studies.

In the end, as suggested by his/her teacher, Yamada promised to help the classmate with math for the upcoming math test. They set up they would meet on Sunday morning to study, as the test was on Monday. The classmate showed great appreciation.

On Sunday, Yamada spent the day doing groceries and cleaning the house with his/her mother. At night, when (s)he decided to review for the test, (s)he realized (s)he had forgotten about the promise to the classmate.

On the next day, they had the math test. A week later, Yamada heard the classmate had failed the test.

5. Hayashi lived alone in his/her apartment. Hayashi's mother lived with his/her sick grandfather at the other side of the city. One day, Hayashi's mother called. She said she had to do a night shift, so she asked if Hayashi could spend the night with the grandfather and make sure he would take his medicine.

Positive desires. Hayashi really wanted to take care of the grandfather. (S)he liked him since childhood and wished he would recover soon.

Neutral desires. Hayashi didn't specifically want to take care of the grandfather but didn't mind it either. (S)he had a distant relationship with the grandfather and was overall neutral towards him.

Negative desires. Hayashi really did not want to take care of the grandfather. (S)he disliked him since childhood and wished he wouldn't recover soon.

[However], in the end, Hayashi promised to take care of the grandfather.

That day, Hayashi was exhausted after work and fell asleep immediately after getting to bed. The next day, Hayashi woke up realizing (s)he had forgotten about the promise to take care of the grandfather.

Later, (s)he found out that, as the grandfather had not taken the medication properly the night before, his condition had worsened.

6. Hirata's friend was traveling abroad. His/her flight was scheduled to depart at 5am. As Hirata had a car, Hirata's friend asked Hirata to drop him/her off at the airport around 3am, since public services did not work at that time.

Positive desires. Hirata really wanted to drive the friend to the airport. (S)he thought friends should help each other and wished the friend would have a pleasant journey.

Neutral desires. Hirata didn't particularly want to drive the friend to the airport but didn't mind doing it either. (S)he felt neutral about the friend's request.

Negative desires. Hirata really did not want to drive the friend to the airport. (S)he thought the friend was inconvenient and wished the friend would have a hard time in his/her journey.

[However], in the end, Hirata promised to drop the friend off at the scheduled time.
On the night before the flight, Hirata was suddenly burdened with lot of work at the office. (S)he worked until late and fell asleep on the sofa after returning home. The next morning, (s)he realized (s)he had forgotten about driving the friend to the airport.
Hirata later found out the friend had lost his/her flight.

Note: Japanese language does not require gender-specific pronouns, therefore the actual Japanese scenarios did not determine the gender of the agent.