Moral decisions from hypothetical dilemmas to the real world:

How do people make moral decisions when faced with multiple probabilistic alternatives (under uncertainty)?

By

Siyuan Yin

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Psychology

December 16, 2017

Nashville, Tennessee

Approved:

Jennifer Sue Trueblood, Ph.D.

David H. Zald, Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Few would disagree that moral decisions are often complicated regarding both decision makers and moral situations embedded in the real world. The complexity of moral decisions revolves around uncertainty and ignorance of the external world and the ambiguity of, or even ambivalence towards, our internal preferences. The uncertainty and ignorance of the external world are of a twofold nature: the probabilities of outcomes of each option are not always available (e.g. how likely will your donation help children from preventable diseases of Kowsar in Iran), and the relationship between alternatives and their outcomes under certain circumstances is difficult to evaluate (e.g., the unpredictability of the terrorist attack makes it difficult to decide whether to improve military training and weapon renovation or to enhance detection of potential violent extremist threats). On the other hand, the ambiguity of our internal preferences (e.g., a doctor treating a terminally ill patient who is suffering from unbearable pain may struggle to decide between honoring the Hippocratic Oath or honor the patient's request to prescribe medication to end his life[1]) manifests itself in the inconsistency of individual choices with the same set of alternatives under seemingly identical situations. This inconsistent preference may be influenced by factors such as mental states, learning, and the environment. Given both uncertainty of the external state and unsure preference of our internal state, *moral flexibility* are prevalent in moral decisions. Moral flexibility is defined as, "people are strongly motivated to adhere to and affirm their moral beliefs in their judgments and choices—they really want to get it right, they really want to do the right thing—but context strongly influences which moral beliefs are brought to bear in a given (Bartels, Bauman, Cushman, Pizarro, & McGraw, 2014).

Consider a real-life example, in which the purpose of punishment represents a moral dilemma for a legal decision. An Ohio teenager, who ditched thirty miles of cab fare, was not automatically given the standard sixty-day jail sentence for juvenile delinquents (Moran, 2015). Instead, the judge offered her two options: one was to serve sixty days in jail (embodied punishment as a means of retribution for wrongdoing), and the other was to walk the thirty miles within forty-eight hours (embodied punishment as a means of rehabilitation). The first question that underlies this dilemma is, do people deserve punishments for wrongdoing, or should punishments deter people from committing crimes? Some may believe people should be

[1] This example is modified from Bartels et al. (2014).

punished merely for the nature of their misconduct; in this case, we would send the woman to jail. Others may argue that "punishment" provides an opportunity for the individual to reconsider his/her choices, for example, walking 30 miles in 48 hours in this case would give the young woman a second chance. These two considerations underscore the flexibility of moral judgments which has been often observed in studies of moral psychology, as people may have conflicting thoughts derived from different moral principles in mind; this is the complexity of our internal world.

The second question of this dilemma is that the judge needs to make judgments under uncertainty, meaning the subsequent influence of each option on this young woman is unknown. How likely would she resume her normal life after serving in jail? Or how likely would she continue harming society if she were "penalized" in such a non-traditional way? From the judge's perspective, sending the woman to jail would affect the rest of her life. That might not be morally acceptable to the judge. Providing her with an alternative might influence her less negatively. Unsurprisingly, the culprit chose the latter. The judge might still be susceptible to regret if he later found that she continued harming society. But he would have to make a decision without future information becoming available to predict whether this person would commit crimes in the future or not. The uncertainty the judge faces is from the external world.

Current literature in moral judgments mainly focuses on hypothetical static and deterministic dilemmas, which may not be applicable to real world moral judgments and may not reveal the real underpinnings of moral judgments. For instance, the trolley problems (Foot, 2002; Thomson, 1985) have been widely discussed and studied in both philosophy and psychology to test theories of moral judgments in harm-based moral dilemmas. The basic two trolley scenarios are the bystander and footbridge situations (see other variants in Hauser, Cushman, Young, Jin, & Mikhail, 2007), in which a trolley is speeding out of control toward five railway workers. In the bystander scenario, a bystander can flip a switch to divert the trolley onto a side track, which will save the five workers but kill one person on the side track. In the footbridge scenario, a bystander on a footbridge over the trolley track can push a sufficiently heavy man next to him onto the track to stop the trolley, so that the large man will be killed and the five others will be saved. The consequences of actions under these two conditions are deterministic, namely, pushing the large man will definitely stop the train. The most common decision pattern is that people approve of flipping the switch to divert the trolley but disapprove of pushing the large man to stop the

trolley to save five persons. Research has reported that when people are asked to make moral judgments under the footbridge condition, they still tend to have counterfactual alternatives to reality (e.g., people think "if only…", or "what if…" and imagine how the past could have been different), even when they are told that the consequences of chosen actions will surely happen (Byrne, 2016). The reason that people refuse to push the large man onto the sidetrack may be that they think it would be impossible for the large man to stop the trolley, suggesting that people do take the probabilistic nature of moral scenarios into account. Also, people may believe that there are often more than two alternatives or solutions to choose from in any given single situation. In the bystander scenario, people may think that the one worker on the side track would escape before being hit by the trolley, which makes the switch action more justifiable than doing nothing. In light of the above issues, it would be necessary to explore how people arrive at certain moral judgments when faced with probabilistic and multiple alternatives (under uncertainty), in order to further understand the processes of making moral judgments.

Given the multifaceted nature of moral judgments, I do not claim that decision theories from non-moral domains are sufficient to explain or describe the actual processes underlying moral decisions. Instead, this article attempts to inspire some future studies on real-world moral decisions, especially under risk and with more than two options. In the first three chapters, I review experimental findings with hypothetical static and deterministic moral dilemmas, including moral principles and rules as normative and descriptive moral theories; factors that influence moral decisions but cannot be formulated as rules that we can use under different moral scenarios; and two representative examples of models of moral decisions which may reveal descriptive processes of moral decisions. The fourth chapter focuses on the relevant non-moral decision theories, with the goal of providing the fundamentals of decision theories and seeking commonalities between moral and non-moral decisions. Possible future directions will be discussed in the fifth chapter.

CHAPTER I

Moral rules and principles

Moral judgments are consistent with moral rules and principles including philosophical principles and legal distinctions. Moral principles are believed to be rigid, like the unambiguous prohibition in the Bible, "Thou shalt not kill", and like the law with which we must comply with, or otherwise be punished (Bartels et al., 2014). These moral principles regulate people's conduct and tend to prescribe norms of acceptable behavior, and also provide guidance for people to judge blameworthiness and punishment as a third party. I will refer to these moral principles in this article as *codifiable*[2] principles, as they can be explicitly articulated and directly applied to different situations. In the long-standing rationalist tradition, humans are believed to be capable of rational reasoning. Like mathematics and logic that provide us with axioms and principles, both psychological and philosophical literature suggest that we also rely on moral rules and principles to make moral judgments, though people may yield different judgments under the same moral situations. Codifiable principles can be derived from principle-based ethical theories such as deontology and utilitarianism; or from observed behavioral patterns such as preference of indirect harm and omission. These principles attempt to reach a conclusion with axiomatic or at least explicit principles that are situation-independent and universally applicable via appropriate mental computation (or moral reasoning). This section will focus on two extensively studied moral principles—deontology and utilitarianism, morally relevant norms, and causal structure theory that is concerned with the relations between acts and outcomes.

Deontology refers to normative moral rules of duty and obligation on what to do and what not to do. These rules concern acts out of duty, but not of our own inclination or of the value of consequences (Kant, 1996). That is, the justification for action should come from the "reason" alone, not from the motive of self-benefit or of the best consequences. Many deontological rules, for instance, prohibit actions that cause harm, regardless of the value of the consequences. Kant is the most representative among modern deontologists. Kant's formula of the universal law is "I

[2] Codifiable comes from the verb "codify", which is defined in the Oxford dictionary "to arrange (laws or rules) into a systematic code".

ought never to proceed except in such a way *that I could also will that my maxim should become a universal law.*" (Kant, 1996) That said, we should only act according to a maxim which can be generalized to a universal law is often tested whether its application to certain situation is generalizable to other situations. For example, taking an emergency exit to save time for an important meeting would not be justified, when we generalize this maxim (i.e., take emergency exits to save some time) to a universal law (i.e., everyone can take emergency exits whenever needed). If everyone took the emergency exit, the emergency exit would not exist anymore and would be just a normal exit. Following this reasoning, we should not take the emergency exit merely to save time for a meeting. Kant's second formula states: "So *act that you use humanity, in your own person as well as in the person of any other, always at the same time as an end, never merely as a means.*" (Kant, 1996) That said, the justified response to the value of humanity is respect; respect for humanity will allow and sometimes require certain action or inaction that does not necessarily achieve the "best" consequences. For example, in the footbridge trolley dilemma, the principle of never treating others merely as a means to an end does not permit pushing the large man, as the large man would be utilized as a means to stop the trolley, just like a heavy object.

The alternative to deontology is utilitarianism (Smart & Williams, 1973), known as the principle of the greatest good, which states that the best decision is that which maximizes the values or utilities of consequences. That is, whether acts are morally forbidden, permissible, or compulsory would be merely determined by whether such acts maximize aggregate welfare. Utilitarianism may approve acts prohibited by deontological demands. In both basic trolley dilemmas, utilitarians would choose to prevent the trolley from hitting the five workers either by flipping a switch to divert the trolley or by pushing the large man off the bridge onto the sidetrack to stop the trolley. In utilitarianism, saving five persons at the cost of one has greater value than keeping one person alive. This principle can be applied to other variants of the trolley dilemmas in which people are asked to choose between saving more people at the cost of a few lives and letting more people die without harming the few innocents—regardless of the approach to achieve the greatest good. In practice, however, people are more likely to flip the switch in the bystander scenario but refuse to push the large man in the footbridge scenario. The results indicate the complexity of moral judgments. It seems that optimizing the number of lives is irreconcilable with the justification of sacrificing an innocent person. Also, it is worth noting that

utilitarianism presupposes that all possible consequences are achievable and can be computed and balanced/compared to find the optimal solution, though not every utilitarianism requires thorough computations under all situations[3]. The online computation algorithm required by the welfare maximizing principle in utilitarianism does not take our limited cognitive capacities into account.

Besides deontology and utilitarianism, social norms regulate actions that are required, permissible, or forbidden independently of any legal or social intuition. These norms are considered culturally universal and ubiquitous in the lives of people, though the contents of norms can be sometimes strikingly diverse: people exhibit reliable compliance with the norms of their group through internalization or intrinsic motivation (Sripada & Stich, 2006). Social norms are often applied to determine whether an act is blameworthy and deserves punishment (retributive justice), and to determine the fairness of distribution (or egalitarian distribution) (Darley & Shultz, 1990; Fehr & Fischbacher, 2004). Some social norms are believed closely related to social relationship. The relationship regulation theory postulates the existence of four fundamental and distinct moral motives: unity (support the integrity of in-groups), hierarchy (respect rank in social groups), equality (balanced reciprocity), and proportionality (reward and punishment are proportionate to merit) (Rai & Fiske, 2011). According to relationship regulation theory, whether any action is judged morally correct or incorrect depends on the employed moral motives and relevant social relationships constructs. Most American people would hold some extent of moral obligations to friends and kin, but much fewer would show comparable obligations to strangers in developing nations (Prinz, 2006).

Other codifiable principles are based on causal structure theory, which focuses on people's interpretation of the relationship between actions and consequences, which often occur subsequently and may or may not be directly caused *by* the actions. People's moral decisions, particularly decisions on punishment and blameworthiness, have been found to rely on at least one element of the causal structure (e.g., the direct causal relationship between an act and an outcome). Also, some biases in moral decisions have been observed in psychological experiments within the causal structure framework (Baron, 2008; Cushman & Young, 2011; Iliev,

[3] Rule-utilitarianism, whether the rightness of taking particular actions depends on its conformity with certain rules that lead to the greatest (Smart & Williams, 1973), does not require computation for each individual case. But how people should act when faced with conflicts between rules is under debate.

Sachdeva, & Medin, 2012; Waldmann & Dieterich, 2007), which will be also included in the following.

The Intention principle, also known as the Doctrine of the Double Effect (DDE) Harm can be intended as a means or as an unforeseen side-effect of acts. The double effect refers to the effect that an action might yield two opposing judgments: people often judge intentional harm to achieve a greater good (the intended goal effect) as impermissible, and unintentional harm (a foreseen but unintended side effect) as permissible (Cushman, 2016; Cushman, Young, & Hauser, 2006; Foot, 2002; Hauser et al., 2007). This bias is known as the intention principle or DDE. DDE is often related to the evaluation of intention, which often requires the ability of perspective taking, or theory of mind (to put yourself in another person's shoes). Intention evaluation serves as a critical factor that affects moral judgment. One possible way to clarify the situation is to implement the principle of informed consent that can ward off treating others as a mere instrument. In the footbridge, if the bystander chooses to push the large man off the bridge without informed consent, the harm to the large man is not likely permitted according to DDE. There is not enough information for us to judge the intention of the bystander, namely whether the harm is intended or only the side-effect.

DDE can also provide an alternative explanation for people's indirect action bias: action that directly causes harm will be judged more morally impermissible than action that indirectly causes the same harm. "Pushing" is often judged more morally impermissible than "switching" in the trolley problem. That is, the outcomes caused by direct action are often considered as intentionally (e.g., using the large man as a mere means), but outcomes caused by indirect action are often considered as a foreseen side-effect (e.g., flipping the switch to divert the trolley that causes the death of one person as a side-effect). Nevertheless, some researchers argue DDE itself is not sufficient to exert a categorical boundary on moral judgment, namely, people cannot determine whether an action is a moral transgression only based on DDE (Cushman, 2016). However, an online study involving several thousand participants replicated DDE across demographical variations (Hauser et al., 2007).

Intention's central role in moral judgments can be reflected in two special phenomena observed in moral psychology. One is the *Knobe effect*, or the "Side-Effect Effect" that refers to the asymmetry in responses (judged as either intentional action or only a side effect) when people judge two similar scenarios in which the side effect of the action differs (Knobe, 2003;

2004). That said, moral evaluations can influence mental state ascription, in particular, whether a described behavior has been conducted intentionally. The common case used to study the Knobe effect is the chairman and environment case:

The chairman of the board of a company has decided to implement a new program. The vice-president reported to the chairman:

(1) that the program will make a lot of money for his company; and

(2) that the program will also harm the environment.

The chairman of the board answered, "I don't care at all about the environment. I just want to make as much profit as I can. Let's start the new program." They started the program. As it happened, the environment was harmed.

People judge the chairman intended to harm the environment. However, if "harm" is replaced by "protect", people judge the chairman protects the environment only as a side effect of launching the new program. One explanation for this result is that behavior of conforming to norms (moral or otherwise) is less informative about underlying mental states than is behavior of violating norms (Uttich & Lombrozo, 2010). It implicates the important role of theory of mind in moral judgments. We are not sure about whether the behavior that is consistent with the norms is indeed performed intentionally (the chairman launched the program that protects the environment), so the protection is more likely to be judged as a side-effect than an intended outcome. In the case where the chairman launched the program that harms the environment, we become more confident that behavior of violating the norms is more likely done intentionally, so the harm is more likely to be judged as an intended outcome.

The other phenomenon is called *moral luck*, which refers to a situation in which people who choose to do the right thing accidentally cause a bad outcome and are assigned some degree of punishment (Martin & Cushman, 2016; Young et al., 2010). Many moral judgments are asymmetrical, in the sense that the "unlucky agent" would be judged more morally blameworthy. A young mother would be judged blameworthy or would even face legal charges if she left her 2-year-old son unattended in the bathroom to pick up a phone call and then found that he was face down in the tub. An explanation for the asymmetry is that the unlucky agent has a false belief (e.g., her son was safe outside the tub), so she is judged as morally blameworthy because of having less justified beliefs (Young et al., 2010). This result also suggests that assessment of the agent's mental state, or theory of mind, can influence people's judgments. Martin and

Cushman (2016) further dissociate an agent's intent to cause harm, their causal role in the harm, and the degree of control the agent had over their behavior. They found that people who choose to do the right thing but accidentally cause a bad outcome receive more moral condemnation than people who are forced to do the "right" thing but cause a bad outcome. The ironic effect suggested that the control of the agent over her behavior had a unique impact on punishment which depends on the attribution of causal responsibility to the agent (Martin & Cushman, 2016).

Omission bias refers to the tendency to judge acts that are harmful as worse than omissions that are equally or even more harmful, given that intention, motives, and consequences are held constant in the context of moral dilemmas (Baron & Ritov, 2004; 2009; Cushman et al., 2006; Royzman & Baron, 2002; Spranca, Minsk, & Baron, 1991). "Harm is an act, but failing to prevent harm is an omission" (Royzman & Baron, 2002). Spranca et al. (1991) found that a few people were surprisingly willing to accept more harm to avoid action. They provided potential explanation to the omission bias that at the moral level, people may "feel righteous by abstaining from sins of commission", while at the personal decision level, people can "avoid blaming themselves for their own misfortunes that they could have avoided through action".

Locus of intervention refers to the notion/idea people's attention may be shifted toward different aspects in moral scenarios so that different moral judgments are made depending on which aspect has been intervened (Waldmann & Dieterich, 2007; Waldmann & Wiegmann, 2010). The locus of intervention falls under the framework of force dynamic theory (Wolff & Song, 2003). In the bystander scenario of the trolley problem, the *agent* (the trolley) actively brings harm to *patients* (five workers on the track). This scenario is called agent intervention. In contrast, in the footbridge scenario, to push a large man (patient) off the bridge to stop the trolley is actually to redirect harm to the large man who replaces the five workers, which is called patient intervention. The placement of intervention on agents or patients affect people's moral judgments, as intervention may change the salience of elements in one moral dilemma by redirecting the locus of attention (Waldmann & Dieterich, 2007). An attentional focus on the victims manipulated by the intervention (removing threat from five workers by diverting the trolley) at the neglect of one innocent person likely leads more judgments of moral permissibility.

Physical proximity refers to whether direct physical contact with the victim (the single worker on the side track in the bystander scenario, or the large man on the footbridge) is involved to take an action. Studies have shown that people are more unwilling to take actions

which involve direct physical contact (Cushman, Young, & Greene, 2009; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). Greene and colleagues (2009) manipulated three conditions with different degrees of physical proximity in the trolley problems: physical contact, spatial proximity, and person force. They found that the moral acceptability of sacrificing one life to save five involving physical contact was rated lower compared to those harmful actions only involving spatial proximity and person force. One possible explanation is that moral dilemmas involving physical contact engage emotional processing, which prevents us from harming the innocent (see Chapter 2).

As reviewed in this section, codifiable principles seem to provide people with the best opportunity to apply those explicit principles to complicated moral situations. They also have the advantage of reducing the amount of mental effort of cost-benefit calculations in repetitive and similar scenarios. As such, we may be able to calculate the optimal action we ought to take according to certain moral rules. However, these codifiable principles may have limited applications because of their rigidity, as optimization under certain rules cannot justify potential violation of other rules. Lying is frequently considered immoral based on deontological rules, but lying to a Nazi officer searching for Jews would lead people to reevaluate the principle of honesty. Rigid principles, in many situations do not constitute a coherent system for people to follow. Being honest to the Nazi officer might in result deeper self-condemnation at the death of Jews. This is one of the reasons that moral judgments seem puzzling to us, i.e. people's moral judgments are not always consistent with one or a restricted set of rules. Also, one may argue that we may apply different principles under different circumstances. If that is the case, how certain principles are selected and how the conflicts between different principles are solved would be the next question to be answered. Precise computations and comparisons based on codifiable principles, in theory, enable people to arrive at "optimal" decisions, but in many situations part of us leaves still unsatisfied, *simply* because we do not feel right about those decisions. It is hence beneficial and sometimes necessary to judge beyond these principles, such as factors that influence our moral judgments but cannot be explicitly articulated or generalized across different moral contexts, which will be discussed in the next chapter.

CHAPTER II

Flexibility of moral judgments

In many situations people's judgments cannot be fully explained by codifiable principles alone. Rather, there is a great deal of nuance in moral judgments across seemingly similar scenarios. In this regard, two aspects of mental processes could contribute to people's moral judgments, so that they become flexible. First, people's sympathetic or aversive emotions, and many other types of emotions may influence their judgments. In the Ohio judge case, it would be reasonable for the judge to send the young woman to jail according to the law, in other words, to render a decision based on codifiable principles; and if he had not offered the alternative punishment, few would criticize his "harshness". However, the alternative of walking for thirty miles might strike some people as more acceptable in this situation. The record of being sentenced to sixty days in jail may affect the woman's life more severely than the alternative. In this situation, people may feel sympathy for the young woman, so they might be willing to accept the other decision as "punishment". Second, people under certain circumstances are unable to explicitly elaborate comprehensive explanations for their moral judgments, called "moral dumbfounding" (Cushman et al., 2006; Dwyer, 2009; Hauser et al., 2007; Schnall, Haidt, Clore, & Jordan, 2008; Wheatley & Haidt, 2005). Consider cases in music and language; prominent musicians play their own interpretation of pieces without explicit reasons or rules we can apply to other even similar pieces; or we insert words into conversations that are not grammatically necessary, but the conversation feels more natural with these seemingly "unnecessary" words. In both cases, agents involved naturally know how and what to do without clearly prescribed rules that can be adopted and generalized to guide future actions. For simplicity, I adopt the term *uncodifiable*[4] principles from McDowell's Virtue Theory (1998) to

[4] "This picture fits only if the virtuous person's view about how, in general, one should behave are susceptible of codification, in principles apt for serving as major premises in syllogism of the sort envisaged. But to an unprejudiced eye it should seem quite implausible that any reasonable adult moral outlook admits of any such codification. As Aristotle consistently says, the best generalizations about how one should behave hold *only* for the most part. If one attempted to reduce one's conception of what virtue requires to a set of rules, then, however subtle and thoughtful one was in drawing up the code,

refer to those factors affecting our moral decisions that cannot be explicitly articulated, or that provide reasons to justify people's judgments even through *post hoc* reasoning (i.e., people may justify their moral judgments after the judgments have been made). These are different from those codifiable principles exemplified in the first chapter.

These uncodifiable principles consist of multiple factors, like affective intuitions, political beliefs, personality, inter-personal relations and culture. Other factors, such as personal past experience (e.g., educational background), moral characters (e.g., courageousness), social convention, and mental state when making moral judgments, also play important roles in shaping people's moral judgments but will not be discussed in this article.

Affective intuitions refer to a process similar to perception, through which people make moral decisions, rather a process of ratiocination and reflection. Social intuitionists asserted that "incommensurable moralities are on top of a foundation of shared intuitions" (Haidt & Joseph, 2004). That is, moral intuitions come first and directly cause moral judgment (Haidt, 2001). The implications from studies about the relationship between affective intuitions and moral judgments are (i) affective intuitions often lead to moralization; (ii) perceived violation of moral principles or norms often evoke morally relevant negative emotions, such as guilt, shame, anger, contempt, or disgust (Huebner, Dwyer, & Hauser, 2008).

Some moral judgments are believed to be rooted in evolved, automatic, consciously inaccessible intuitions, in which moral reasoning is driven by moral intuitions "just as surely as a dog wags its tail" (Greene, 2007a; Haidt, 2001). Moral dumbfounding has been observed in hypnotizable people (Wheatley & Haidt, 2005). These participants were hypnotized to feel a flash of disgust whenever they read an arbitrary word, for example, "often". The results showed that moral judgments were made more severe with the hypnotized word related to the presence of a flash of disgust. Similarly, the severity of moral judgments made by participants, who actually experienced feelings of disgust during the experiment or recalled a physically disgusting experience, was greater than those of the control (Schnall et al., 2008). In both experiments, participants could not justify their moral judgments based on reasoning. One participant wrote at the end of study: "When 'often' appeared I felt confused in my head, yet there was turmoil in my

---

cases would inevitably turn up in which a mechanical application of the rules would strike one as wrong…" (McDowell, 1998) (p.57-58).

stomach. It was as if something was telling me that there was a problem with the story yet I didn't know why" (Wheatley & Haidt, 2005).

Not only emotion invoked by given moral dilemmas, but also emotions induced by the environment have an influence on moral judgment. If the two emotions have opposite valence, emotions induced by the environment likely reduce the effect of emotion related to dilemmas on ultimate judgment. Valdesolo & DeSteno (2006) presented either comedic or affectively neutral video clips to two groups of participants. They found that inducing positive contextual emotion (by watch comedic video clips) increased morally appropriate responses (i.e., making people's judgments of harmful actions more utilitarian in the push case), which indicates that manipulation of contextual emotion shapes moral judgment (Valdesolo & DeSteno, 2006).

Notwithstanding the well-recognized role of emotion in moral psychology, the refined source of the emotional influence throughout the whole processing of moral judgment, involving context perception, interpretation of moral dilemmas and questions, moral judgment production, and actual responses or moral actions, has yet to be further elaborated. What needs to be clarified is whether emotional mechanisms "constitute moral concepts" or "merely stand in an important causal relationship" to them (Huebner et al., 2008). For instance, the affect triggered in moral judgments has been reported differentiable, as people are found to have different aversions toward different attributes of one alternative (Miller & Cushman, 2013), e.g. action-based aversion and outcome-based aversion.

Political beliefs can influence people's interpretation of moral foundations, which consist of harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, and purity/sanctity (Graham, Haidt, & Nosek, 2009; Haidt, 2007). Among the five components, harm/care and fairness/reciprocity are more related to individual welfare; the other three, called binding foundations, focus more on social cohesiveness and social order, respect of authority, duty, and obedience. Moral foundations are proximate to being innate in the sense that they are also subject to evolution over people's lifetime. People who are more politically conservative are more likely to consider the five moral foundations equally. In contrast, people who are more liberal put more weight on the foundations related to individual welfare and rights (Haidt, 2013), and it seems more acceptable for them to make trade-offs on the binding foundations.

Personality Studies pertaining to personality and characteristics related to morality with psychopathic and other clinical groups have been explored "how differences in the propensity to

rely on intuitive reactions affect judgments" (Bartels, 2008). In the study, people who scored high in psychopathy and Machiavellianism would devote higher endorsement to utilitarian judgments in moral issues (Bartels & Pizarro, 2011). Such results raised the concern about current deontology-utilitarian categorization of moral traits, as to whether or to what extent the endorsement for more utilitarian choices comes genuinely from people's belief of maximizing the overall welfare rather than ascribing some deficits in certain psychological processes.

Inter-personal relations can also influence moral judgments, as we do not live alone but with other people. Some moral principles (e.g., fairness/reciprocity, keeping promises) are embodied in interpersonal relations. For instance, people expect the fairness of distributions within groups (having at least two persons), meaning that people tend to punish those who are overrewarded and reward those who are underrewarded. Moreover, studies found that people were even willing to sacrifice their own interest for the sake of equity. In the first part of a dictator game (Kahneman, Knetsch, & Thaler, 1986), a subject (dictator) was asked to divide $20 between himself and another subject (receiver) who could not reject the allocation. They did not know who were their receivers. The dictator could keep $18 to himself and gave $2 to the receiver, or $10 to each. Among all subjects, 76% divided the $20 evenly. Only a portion of the participants actually received the money. In the second part of the game, with subjects who were not paid in the first part, a dictator would be asked to divide money between himself and another two receivers (who were dictators in the first part). The dictator was told about his two receivers, one who had divided $20 in the first part evenly (E: $10 to each) and the other who divided $20 unevenly (U: $18 to himself and $2 to the other). The dictator could decide between option A ($5 to himself, $5 to E, and $0 to U) and option B ($6 to himself, $0 to E, and $6 to U). Again, 74% chose option A, meaning that the majority chose to sacrifice in order to punish the U. Therefore, people expect others in the same group to conform with fairness/reciprocity, or equity and are willing to punish violators even at the cost of self-interest.

Culture has been believed to play an indispensable role in shaping moral decisions and the development of morality (Bersoff & Miller, 1993; Miller & Bersoff, 1992). Two groups of people, Indians and Americans, were first asked to rate the undesirability of a justice breach (e.g., Fred takes another man's train ticket without the man's permission) and an interpersonal breach (e.g., Jim does not deliver the wedding rings to his best friend's wedding). They were then presented situations with both justice and interpersonal conflicts, and asked to choose resolving

either justice or interpersonal conflicts at the cost of the other. The results showed that Indians more often resolved the conflicting situations in favor of the interpersonal options than did Americans (i.e., Indians are more likely to choose to meet the interpersonal demands). Furthermore, some people are found to hold "protected values" (values that are protected from others, e.g., value of natural resources) which lead to differences in moral judgments (Baron & Spranca, 1997). People may have or be sensitive to different protected values. For instance, some people find abortion or destruction of natural resources absolutely unacceptable, as either lives or natural resources are not even comparable to other values. In studies, people with certain protected values showed consistently strong opposition to some actions in moral scenarios despite benefits (Baron & Spranca, 1997).

Some biases related to the framework of uncodifiable principles have been studied, which may derive from counterfactual thoughts (Byrne, 2016), and may be due to framing effects, order effects, priming effect, attention allocation, and context effects. These biases have also been examined in non-moral decision-making domains. The commonalities of decision biases may reveal certain similarities of the decision-making processes of both moral and non-moral domains, as the above biases have been observed in moral judgments. This examination of these shared biases is important. To understand moral decisions thoroughly, we need to know both the basic theoretical structure constituting the set of moral principles and the perceptual factors determining how we perceive specific moral contexts (e.g., the semantics and logical structure in moral dilemmas).

Counterfactual thoughts are spontaneously created by people to compare reality with counterfactual alternatives. They suggest how the past could have been different. The process of computing counterfactuals is believed primarily automatic (Byrne, 2016). The comparison between reality and counterfactual alternatives has been considered to serve as "a powerful social glue" supporting moral judgments, in particular, on judgments about intentions, blameworthiness and punishments (Byrne, 2016). In the trolley problem, for example, people may have counterfactuals to think of what they could have done to avoid the worst situation (killing the innocent), so they judge that it is not morally permissible to push the large man off the bridge onto the sidetrack. Also, consider the chairman and environment example (see Knobe effect in Chapter 1). In one case, the chairman is judged to intend harm toward the environment, whereas in another case the harm toward the environment is seen as a side-effect of the

chairman's decision. People may think that the chairman could have chosen to stop launching the new program to avoid harming the environment; whereas in the protection situation people may think that there is no dilemma for the chairman, because launching the program could help protect the environment at the same time make a profit. Thus, it seems more likely that the chairman intended to harm the environment (Byrne, 2016; Knobe, 2003).

Framing effects refer to preference reversals for the same situations which are framed in different ways (Tversky & Kahneman, 1981). Petrinovich & Oneill (1996) investigated the framing effects in moral judgments. In their study half of the moral questions were worded as whatever actions were or were not taken would kill some and not others, and the other half were worded as the actions would save some and not others. Note that the outcomes of both halves were identical. The results showed that the actions in questions with "saving" wording were more often accepted than those with "killing" wording (Petrinovich & ONeill, 1996). Not only the wording in moral scenarios but also wording in moral questions have been found to influence moral judgments. In one study, participants were randomly assigned to respond with one of the four adjectives (i.e., "wrong", "inappropriate", "forbidden", or "blameworthy") (O'Hara, Sinnott-Armstrong, & Sinnott-Armstrong, 2010). The results indicated that it is more likely for people to judge an act as wrong or inappropriate than forbidden or blameworthy. Thus, the framing effects can affect moral judgments from different aspects of moral contexts.

Order effects refer to the fact that people's moral judgments (Wiegmann, Okan, & Nagel, 2011) and endorsements of moral principles (Schwitzgebel & Cushman, 2012) are subject to the sequential presentation of moral situations. That is, moral judgments on one moral scenario are not independent of the previous scenario and the former may serve as a prime for the latter. Note that the order effects observed in moral judgments are often asymmetrical. Specifically, the action (i.e., flip the switch) of the bystander trolley problem are judged worse (less morally acceptable) when preceded by the push problem (often judged unacceptable) than when judged alone. Wiegmann et al. (2011) exaggerated the order effects by presenting participants five moral scenarios in the order of ratings on agreement with the actions from the most unacceptable (i.e., push dilemma) to the most acceptable (i.e., bystander dilemma) (ascending) or the reverse (descending). They found that the ratings on the bystander trolley problem were much worse when the five scenarios were presented in the ascending order. The order effects are also present

in professional philosophers and people having sufficiently ethical education in philosophy (Schwitzgebel & Cushman, 2012).

Context effects, also called *decoy effects*, refer to the role of context on decisions with multiple alternatives. Context effects occur when a third alternative (i.e., S, A, or C) is added to the original pair of alternatives (i.e., X and Y) (Figure 1A). Three effects have been extensively studied: the similarity (Tversky, 1972), attraction (Huber, Payne, & Puto, 1982), and compromise (Simonson, 1989) effects. Similarity effects: when the third alternative is similar to one of the original pair, the probability of selecting the dissimilar alternative is increased. The third alternative becomes a competitor of X, so the probability of selecting X decreases. Attractive effect: when the third alternative is inferior to X, making X more attractive than before (and also more than Y). The probability of selecting X is increased. Compromise effect: when the third alternative serves as a compromise between X and Y, the third alternative is likely selected.

In moral judgments, only the similarity effect has been observed (Shallow, Iliev, & Medin, 2011). Shallow et al. (2011) used trilemmas (Figure 1B: 1- Push intervention, 2- Switch intervention with two people being threatened, and 5 — Omission, or 1, 4 — Switch intervention with four people being threatened, and 5, henceforth called 125 and 145, respectively) which combined two classic trolley problems, bystander and push scenarios, with varying numbers of lives that can be saved. They found that pushing was disapproved of more in the context with 125 and less in the context with 145 than that in original push trolley problem. The result is consistent with similarity effect, where adding a third alternative (a switch intervention) decreases the approval for the similar alternative. This study expands the limit of moral research with binary choices, though other context effects have yet to be examined in moral judgment. To understand context effects in moral judgments, therefore, will extrapolate findings from binary moral dilemmas to broader contexts with multiple alternatives and potentially with multiple attributes.

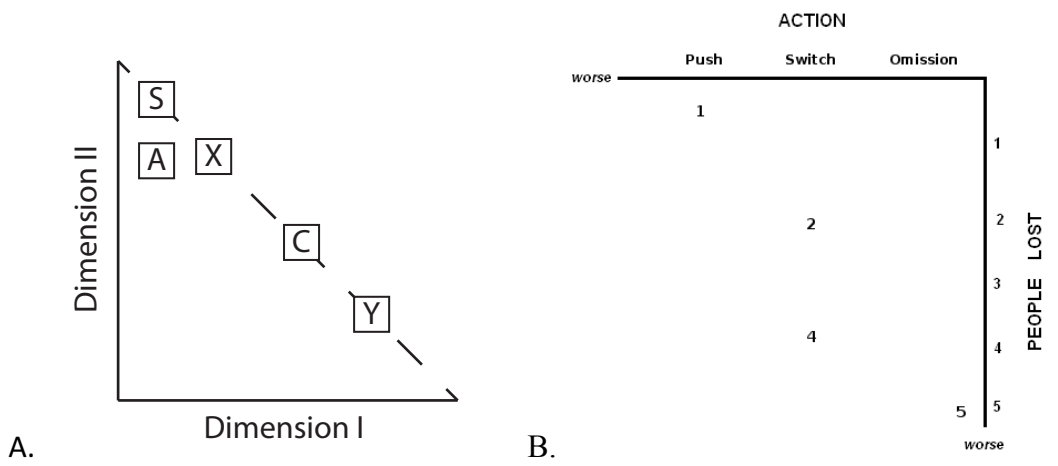

Figure 1. Context effects. A. Consider X and Y are two alternatives (e.g., cars). Neither car is dominating the other, as X has higher value on dimension II (e.g., greater quality) but lower value on dimension I (e.g., expense) than Y (e.g., poorer quality but inexpensive). Context effects arise when we add a third car to choose. Similarity (S) effect: the third car that is similar to but different from (equally attractive as) X increases the probability of Y being chosen. Attractive (A) effect: the third car that is inferior to X increases the probability of X being chosen. Compromise (C) effect: the third car that serves as a compromise between X and Y is more likely being chosen than X and Y. Specifically, C is more economical than X and has better quality than Y. Examples are from (Trueblood, Brown, Heathcote, & Busemeyer, 2013). B. Context effects in moral dilemmas. 1: Push intervention: The five can be saved by pushing a single person off a bridge (but 1 person dies). 2 and 4: Switch intervention: The five can be saved by redirecting the trolley onto a different track (but 2 or 4 people die). 5: Omission: Doing nothing. No deaths are directly caused but 5 people die. Adopted from Shallow et al. (2011).

Attention allocation, or momentary attentional shift, has been observed to influence moral judgments. Manipulation of people's attention allocation can lead to different moral judgments in both static (Bartels, 2008) and dynamic (Pärnamets et al., 2015) manners. For example, Bartels (2008) made either the outcomes or actions under moral situations salient by adding parentheses to these aspects. The salience of outcome and action in the presented moral dilemmas was found to lead to different moral judgments. Pärnamets et al. (2015) manipulated the fixation duration on a randomly predertermined one of two alternatives (e.g., "sometimes justifiable", or "never justifiable") during participants making judgments on a morally related statement (e.g., Murder is sometimes justifiable). Surprisingly, people's judgments were strongly biased toward the alternative they fixated on for a longer period of time. This finding seems to challenge the view that moral decisions are purely products of specific morally motivated intuitive and principle-governed processes.

From all flexibilities of moral judgments discussed in this chapter, moral judgments have their uniqueness, but it would be difficult to oppose the continuum between the moral and

general decision-making domains. Moral principles are rigid; yet moral judgments are subject to diverse uncodifiable factors and share commonalities in decision biases with other choice domains. A parsimonious examination on specific moral theories, moral cognition, and their relationship with non-moral decision making will be critical to fully understand the precise computational and algorithmic real-time underpinnings of moral choices.

CHAPTER III

Models of moral judgments

The puzzle whether the fundamental processing of moral judgment is more akin to "the emotional dog that wags its rational tail" (moral intuition directly determines moral judgments) (Haidt, 2001) or to "the rational dog that wags its emotional tail" (moral reasoning directly determines moral judgments) (Pizarro & Bloom, 2003) has been long lingering in the realm of moral psychology. Many attempts have been made to consider both codifiable principles (rigidity) and context-dependent factors that influence people's moral judgments but cannot be explicitly codified into a set of moral rules (flexibility) with varying degrees of emphasis on each, for the purpose of revealing the underlying processes involved. Some of these attempts are embodied in specific moral models, which include: the dual-process system (Białek & De Neys, 2017; Greene, 2007b; 2007a; Greene et al., 2001; 2008) and the multi-system process (Cushman et al., 2009), social intuitionist model (Haidt, 2001; Haidt, Bjorklund, & Murphy, 2000), sentimental rules theory (Nichols, 2002; 2004; Nichols & Mallon, 2006; Royzman, Leeman, & Baron, 2009), universal moral grammar (Mikhail, 2007), reinforcement learning models (Ayars, 2016; Crockett, 2013; Dayan & Berridge, 2014; Dayan & Niv, 2008), moral heuristics (Fleischhut, 2013; Gigerenzer, 2010; Sunstein, 2003; 2005), the deontological coherence theory (Holyoak & Powell, 2016; D. Simon, Stenstrom, & Read, 2015), and construal-level theory (Gong & Medin, 2012; Trope & Liberman, 2010). The majority of these moral models incorporate both moral principle-governed reasoning and uncodifiable factors with the exception of the universal moral grammar theory (Mikhail, 2007). To link non-moral decision-making with moral decisions, dual-process theory and moral heuristics will be discussed in detail in this chapter. These two theories are the most relevant models that have been extensively studied in both domains, though the importance of other models in the moral repertoire can hardly be overlooked.

Dual-process theories

The prominent dual-process theory of moral decisions (Greene, 2007a) proposes a potential reconciliation between deontological and utilitarian judgments with "emotional" (i.e., affective) and "cognitive" (i.e., deliberative) systems, which respectively correspond to the well-received System 1 and System 2 of the existing dual-process system (Table 1) (Kahneman, 2003; Sloman, 1996). That is, at the core of dual-process theory, decisions and judgments are products of at least two distinct psychological processes (Cushman et al., 2009; Greene, 2007a). This section will discuss the characteristics, supporting evidence, and challenges in both moral and non-moral domains.

Greene (2005) argued that two views of moral decisions, deontology and utilitarianism, refer to "*psychological natural kinds*". Deontology has been deemed a rule-based morality. Moral judgments should be made on the basis of moral rules, not on the basis of emotions or self-inclination. Thus, deontology may allow and sometimes require certain actions or inactions that will not produce the best consequences (i.e., ends don't justify the means). The deontological rules seem to be conducted through reasoning. In contrast, utilitarianism guides people to act, aiming to achieve the greatest good, which is usually gauged by the consequences (e.g. to maximize the number of lives that can be saved in the trolley problem regardless of how it is accomplished: to flip a switch to divert the trolley onto an innocent person or to push a large man off the bridge). The utilitarianism appears to be related to emotions, as one measurement of the greatest good include happiness. However, later evidence shows that the functional roles of deontology and utilitarianism represent two dissociable psychological patterns, which are emotional and cognitive, respectively (though not strictly) (compared in Table 1). In other words, people's inconsistent judgments in the bystander and push trolley dilemmas may come from competition between two distinct psychological systems (Cushman et al., 2009; Greene, 2007a). To avoid confusion and ambiguity, I use, instead, "affective" and "deliberative" to refer to these two systems.

Table 1. General features of the two systems[5] (Haidt, 2001)

The affective system	The deliberative system
Fast and effortless	Slow and effortful
Process is unintentional and runs automatically	Process is intentional and controllable
Process is inaccessible: only results enter awareness	Process is consciously accessible and viewable
Does not demand attentional resources	Demands attentional resources, which are limited
Parallel distributed processing	Serial processing
Pattern matching; thought is metaphorical, holistic	Symbol manipulation; thought is truth preserving, analytical
Common to all mammals	Unique to humans over age two and perhaps some language-trained apes
Context dependent	Context independent
Platform dependent (depends on the brain and body that houses it)	Platform independent (the process can be transported to any rule following organism or machine)

The *affective* system is related to our moral intuitions and other uncodifiable factors, which often inform deontological judgments. That is why the pure rationality in deontology is questioned, as the varying extent of engagement of emotional processing in moral dilemmas affects deontological judgment. In the trolley problem, for instance, people often disapprove of pushing the large man onto the track to stop the trolley but approve of diverting the trolley to save five others at the expense of one life. People's discrepant responses in the push and bystander dilemmas can be explained by the differing degrees of emotional involvement. Compared to the bystander dilemma (impersonal condition), the push dilemma is considered intuitively "up close and personal" (personal condition), which is more likely to trigger a prepotent, negative emotional response that drives people to disapprove of the personally harmful action (henceforth they are called moral-personal and impersonal conditions) (Greene et al., 2001). Greene et al. (2001) found that brain areas associated with emotion exhibited greater activities during judgments of moral-personal dilemmas and subjects responded faster in this

[5] Recent studies have suggested the possibility of intuitive utilitarianism in moral reasoning, and "hybrid" dual process models (i.e., both systems have affective and deliberative elements) have been suggested as alternative to purely serial or parallel models (Figure 2).

condition than in moral-impersonal conditions. Also, with the influence of the affective system, judgments regarding punishment can be justified because of the nature of the moral transgression rather than its related negative outcomes, as people sometimes choose to punish others primarily based on retributive justice (i.e., the wrongdoers deserve punishment themselves). A betrayal of trust, for example, often leads to a great deal of outrage, e.g., a babysitter neglects a child or a security guard steals from his employer (Sunstein, 2005). Moreover, people's moral condemnation has been observed in harmless actions (e.g., A brother and sister make love or a family eats its dog after it has been killed accidentally by a car.) (Haidt, 2001). Haidt (2001) found that people may still feel a quick flash of revulsion at the thought of incest, even when they were told that two forms of birth control were used and no harm would befall the siblings.

The *deliberative* system is related to our rational reasoning that can be built on but is not limited to the principle of greatest good from utilitarianism. Utilitarianism is, by its nature, "a matter of balancing competing concerns, taking into account as much information as is practically feasible" (Greene, 2007a). Greene et al. (2001) suggested that moral-impersonal dilemmas require controlled cognitive processes by showing that brain areas associated with working memory were found to be more active under moral-impersonal conditions. Furthermore, performing morally irrelevant cognitive tasks (a concurrent digit-search task) during moral judgments has been found to selectively interfere with utilitarian moral judgments (Greene et al., 2008). Specifically, the cognitive load increased the average response time for utilitarian judgments but did not increase that of non-utilitarian judgments, which implies that these two types of judgments are driven by two different processes.

A direct application of the dual-process theory is that it provides alternative explanations for certain biases in moral decisions. The manipulation of the salience of the outcomes or actions within certain alternatives, for instance, altered moral judgments of people who hold certain protected values (Baron & Spranca, 1997). In light of the dual-process theory, trade-offs usually require the engagement of cognitive processes, but people with protected values tend to experience outrage at the thought of making trade-offs, and deny the need for trade-offs (Baron & Spranca, 1997). Different wording in moral dilemmas (e.g., save or kill) may trigger people's different intuitive responses which then lead to even opposite moral judgments on seemingly identical moral dilemmas (Petrinovich & O'Neill, 1996). Also, people under time pressure made more deontological judgments than those without time pressure (Suter & Hertwig, 2011). This

result indicates that moral judgments can be altered by limiting cognitive control via deliberate time manipulation, as moral judgments may then rely largely on the affective processing under time pressure.

Long-standing criticism of the dual-process theory concerns the assumption that the affective and deliberative systems are at large independent, and the lack of specificity on how the two systems interact if at all. One hypothesis of the interaction between two systems is that the potential balance between the affective system and the deliberative system is determined by individual's self-regulatory dynamic, stress, and development level; exerting self-regulation strategies, or willpower, is essential for execution of "difficult-to-achieve intentions" (Metcalfe & Mischel, 1991). Specifically, we are assumed initially driven by impulses ruled by pleasure principle, and are largely indifferent to reason; it is thus difficult to control our actions and to overcome the impulses without adequate strength of willpower (Metcalfe & Mischel, 1991). Willpower plays a role in inhibiting deontological responses in the affective system and enabling the deliberative system to override the power of deontological intuitions in favor of welfare-maximizing behavior. For instance, people can intervene to deliberatively create or suppress counterfactuals, the processing of which is primarily automatic. In some moral dilemmas that evoke more emotional reactions, like the footbridge trolley problem, people will take longer to imagine what could be done to avoid the worst outcome (Byrne, 2016). This self-regulation scheme can be regarded as a theoretical outline of the dual-process model of moral judgment.

Another hypothesis of the interaction between the affective and deliberative systems is the hybrid dual-process model (Białek & De Neys, 2017), which renders conflicts of intuitions or principles within each system, besides conflicts between two systems. This hypothesis differs from pure serial (i.e., the affective system goes either before in Figure 2A or after the deliberative system) or parallel models (i.e., two systems work at the same time, see Figure 2B), which do not address within-system conflicts.

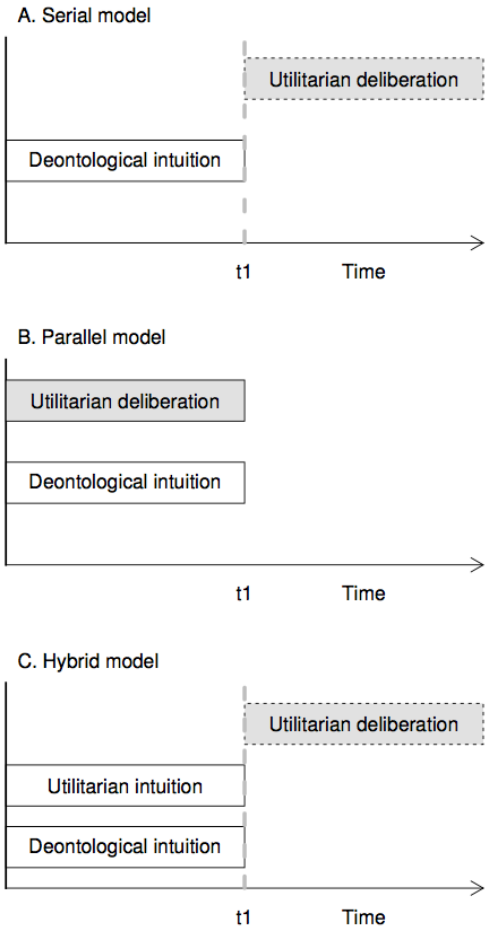

Figure 2. Schematic illustration of three possible views on the interplay between System 1 and System 2 processing in dual process models of moral cognition. Deliberate system 2 processing is represented by gray bars and affective System 1 processing by white bars. The horizontal axis represents the time flow. In the serial model (A) reasoners initially only rely on System 1 processing that will cue an intuitive deontological response. In the parallel model (B) the two systems are both activated from the start. In the hybrid model (C) initial System 1 activation will cue both a deontological and utilitarian intuition. The dashed lines represent the optional nature of System 2 deliberation in the serial and hybrid model that can follow the initial System 1 processing in a later stage (represented as t1). Adopted from Białek & De Neys (2017).

One hybrid dual-process model (Figure 2C) assumes that the affective system is simultaneously generating deontological and utilitarian intuitions under certain classical moral contexts; in this way, the affective and deliberate utilitarianism in the hybrid model are complementary rather than mutually exclusive (Białek & De Neys, 2017). Białek and De Neys (2017) tested this hybrid model by asking participants to evaluate four moral dilemmas while completing a dot memorization load task, which is designed to interfere with the deliberative system processing. The evidence supporting the hybrid model is that people's judgments on the moral acceptability

of sacrificing action were comparable under two conditions. In other words, people exhibited an intuitive sensitivity to the utilitarian aspects of classical moral dilemmas without engaging in demanding deliberation (i.e., under cognitive load condition). Similarly, it is necessary to further specify the general structure within the deliberative system, such as how do various harms and benefits trade off against one another under the welfare-maximizing principle?

Another potential limitation of the empirical evidences supporting the dual-process theory stems from its inference interpretation (Krajbich, Bartling, Hare, & Fehr, 2015). Krajbich et al. (2015) argued that caution should be taken when we utilize response time to infer which choices are intuitive, and to further distinguish mechanisms of the affective and deliberative systems. The potential problem is that the "reverse inference", from behavioral or biological measures to infer mental function, does not take into account other sources of variability in the data (Krajbich et al., 2015). The first source of the variability is due to idiosyncratic individual variability. For example, is giving money to a homeless person an automatic response to help others or a calculated action taken only when the person is truly in need? Different responses time may be due to the difference of altruistic or egoistic personalities. The second source of the variability in the data is due to the choice sets[6] selected by the experimenter, as Krajbich et al. (2015) argued that modifying the choice options appropriately could produce *any* desired RT results. Therefore, using reaction time to make inferences about which choice is from the affective system can be misleading if the choices are now appropriately balanced based on their values; it is also important to note that the presence of a choice-bias alone does not imply a dual-process mechanism (Krajbich et al., 2015).

Dual process theories are not unique for moral decisions, and they have also been proposed in non-moral decision-making domains such as risky decision-making (Evans, 2008; Kahneman, 2003; Sloman, 1996). Roughly speaking, the processes of the deliberative and affective systems have been contrived in a parallel manner (Alós-Ferrer, 2016; Loewenstein, O'Donoghue, &

[6] Consider an example of value-based binary decision making (Krajbich et al., 2015). For each pair of options A and B, suppose the relation between reaction time and the difference in net preference for option A over option B is from a normal distribution with zero mean (i.e., A and B are equally preferred). Subjects make a series of choices, each time between a different A-B pairing. If the experimenter selects an unbalanced set of A-B pairings, reaction time of subjects choosing A or B may be different even on the same group of subjects. When the range of net preference for A over B is not symmetric at zero, for example, more pairings with positive net preference for A over B, the mean reaction time of selecting A is less than the mean reaction time of selecting B.

Bhatia, 2015; Mukherjee, 2010; Van Bavel, Jenny Xiao, & Cunningham, 2012) and in a sequential manner (Guo, Trueblood, & Diederich, 2015). Along the line of parallel mechanisms, decisions from the affective system can be overridden by the deliberative system via sufficient willpower (Loewenstein et al., 2015) or via a central executive either integrating both responses of both system (Mukherjee, 2010) or exclusively selecting the results from one system (Alós-Ferrer, 2016). The rest of this section mainly focuses on these dual-process theories with formal computational models, seeking for potential applications of these theories to moral judgments.

The dual-process model by Loewenstein et al. (2015) assumes that the deliberative system trades off the desirability of actions against the *willpower* effort, an inner exertion of effort required to implement the desired behavior. That is, a person's behavior is "the joint product of a deliberative system that assesses options in a consequentialist fashion and an affective system that encompasses emotions such as anger and fear and motivational states such as hunger, sex, and pain" (Loewenstein et al., 2015). With an assumption that neither of the two systems is completely in charge of behavior, Loewenstein et al. (2015) formalized the "joint product" with two simultaneously operating functions. One is a motivation function from the affective system V_A. The other is a utility function[7] from the deliberative system V_D, and an effort cost h related to willpower. The effort cost depends on two factors: current stock of willpower reserves W, and the number of competing cognitive demands σ. The total value of an option x, denoted by $V(x)$, determines the person's behavior:

$$V(x) \equiv V_D(x) + h(W, \sigma) \times V_A(x, a) \tag{3.1}$$
$$V(A) \equiv V_D(A) + h(W, \sigma) \times V_A(A, a)$$

Here the utility of x is calculated based on utilitarian principles. In the motivation function $V_A(x, a)$, the variable a indicates the intensity of affective motivation. If the willpower is depleted by a recent use (W decreases), or the deliberative system is distracted by irrelevant tasks (σ increases), the deliberative system will have less influence over behavior (exerting willpower becomes more difficult, i.e., h decreases).

[7] Here the utility function can be understood to calculate the value of consequences of actions or choices. For example, in the delay-of-gratification experiment (Metcalfe & Mischel, 1991), a child can get either two marshmallows if she chooses to wait for a longer period of time, or only one marshmallow if she chooses not to wait but get the reward immediately. In this case, waiting leads to two marshmallows, while not waiting leads to one. Typically, the value of two marshmallows is believed greater than one. The calculation of utility of alternatives and related theories will be discussed in Chapter 4.

In the case of altruism and its associated affect sympathy, the deliberative system follows the rationale of moral principles to guide us how we ought to behave, and the affective system depends on the degree of sympathy triggered in certain moral context to determine our motive from pure self-interest to extreme altruism. Suppose each option x of an option set X is a pair of payoff $x = (x_S, x_O)$, where x_S is a payoff for oneself and x_O is a payoff for another person. Thus, $V_D(x) = x_S + \phi x_O$, where ϕ indicates the stable weight the deliberative system puts on the other person's payoff. The more weight on x_O, the greater extent to which the final choice is determined by x_O (thus, the more altruistic behavior the person will exhibit). Similarly, the motivation function $V_A(x, a) = x_S + a x_O$, where a indicates the degree of sympathy the decision maker feels towards the other person. The more sympathy the person feels, the more she would display to the other. With the above components, Equation 3.1 becomes

$$V(x) = x_S + \phi x_O + h(W, \sigma) \times (x_S + a x_O) \tag{3.2}$$

Maximizing Equation 3.2 is equivalent to maximizing

$$\tilde{V}(x) = x_S + \tilde{\phi}(a) x_O, \text{ where } \tilde{\phi}(a) = \frac{\phi + h(W,\sigma) a}{1 + h(W,\sigma)} = a + \frac{\phi - a}{1 + h(W,\sigma)} \tag{3.3}$$

According to Equation 3.3, an increase in $h(W, \sigma)$ (more effort cost due to willpower depletion or distraction by irrelevant tasks) will increase $\tilde{\phi}(a)$ when the affective intensity is high ($a > \phi$), or decrease $\tilde{\phi}(a)$ when the affective intensity is low ($a < \phi$). Also, any factor that increasing the degree of feeling sympathy (increasing a) will increase $\tilde{\phi}(a)$, which will lead to more altruistic behavior (i.e., $\tilde{\phi}(a) x_O$ has more weight on final choices than x_S).

The other two parallel dual-process models assume that a central executive integrates outputs from both the affective and deliberative systems, by weighting the results of two systems (Mukherjee, 2010), or by selecting one process at a rate (Alós-Ferrer, 2016). Mukherjee's dual system model primarily addresses preferences under risk[8], for example, choosing gambles (G).

$$V(G) = (1 - \gamma) V_D(G) + \gamma V_A(G) \tag{3.4}$$

where γ is the weight given to the affective system (the relative extent of involvement of the affective system in risky decision-making). V_D is a linear function of gambles values, denoted by $V_D(x) = kx$ with a constant k. In contrast, V_A is a nonlinear function that is monotonically increasing but with decreasing sensitivity to the magnitude of gamble values, denoted by

[8] Decisions under risk will be discussed in Section 4.1.

$V_A(x) = x^m$ $(m < 1)$, which is one of the characteristics of the affective system, also called diminishing marginal utility. Since Mukherjee's dual system model focuses on risky choice, probabilities in each gamble are also critical in the processing of risky decision-making. Mukherjee (2010) assumed that the deliberative system perceives the objective probabilities without any distortion, namely, $w_D(p) = p$, while the affective system is insensitive to different probabilities, and weights each nonzero probability equally, $w_A = \frac{1}{n}$, where n is the number of nonzero probabilities. Therefore, Equation 3.4 can be reformulated as

$$V(G) = (1 - \gamma)k \sum_i p_i x_i + \gamma \frac{1}{n} \sum_i x_i^m \quad (3.5)$$

Mukherjee's dual system model captures both routes of thinking, though the psychological nature of γ has yet to be further interpreted to determine how to allocate the involvements of both systems.

Alós-Ferrer's dual-process diffusion model consists of a utility decision process (evaluation by calculation) and a heuristic decision process (evaluation by feeling) (Alós-Ferrer, 2016). Different from previous two dual-process models, Alós-Ferrer's model treats each process as a mathematical diffusion process (Ratcliff, 1978). A diffusion process is a stochastic process which continuously describes a random variable (e.g., the motion of a particle) possibly under the influence of noise (e.g., a particle moving under the influence of friction). In a diffusion model, the consideration of each option is regarded as a diffusion process of evidence accumulating towards a prescribed "boundary", and that the final decision is made corresponds to that the diffusion process reaches which boundary. The two processes happen simultaneously and a central executive process select either the heuristic processes with probability Δ $(0 < \Delta < 1)$, or the utility process with probability $1 - \Delta$. The final decision and reaction time are determined by the result of the selected process. With the general assumption that the affective system is faster than the utility system, this model can be applied to predict the relative speeds of responses. Specifically, if the favored response of the heuristic process is inconsistent with that of the utility process (conflict situation), the final response time depends on which process is selected.

Unlike the parallel scheme of dual-process models, Guo et al. (2015) proposed a sequential dual-process model incorporating diffusion processes. This model assumes that: the affective system precedes the deliberative system; the deliberative system evaluates choices with a latency,

whose impact on decisions can vary depending on time pressure or tasks; and there is a switch from the affective system to the deliberative system during the diffusion process. Guo et al. (2015) manipulated the deliberate time which would affect the threshold of evidence accumulation. That is, under time pressure, differences between two boundaries of the diffusion process (in this case, two boundaries represent sure options and gambles) decreases, so a natural prediction is that the affective process may reach its boundary before the switch occurs. The results suggest that framing effects in risky gambling are largely driven by the affective system. Also, if the deliberative process gets into the decision-making process, the favored gamble by the affective system can be overridden by the deliberative system.

Moral decisions, according to theories of dual-process system, that have been believed to result from affective and deliberative processes within decision makers themselves, interactive or dissociable, in a parallel or serial manner, but not to take into account the limitation of our capacity in information processing (bounded rationality) or the impact of the surroundings the decision makers dwell in (ecological rationality) (Fleischhut, 2013; Gigerenzer, 2010). In reality, the amount of information we receive often exceeds what we can actually process and update. Also, the impact of the environment on moral decision has not been fully examined, so moral decisions might be different if decision makers were isolated from the current environment and other people. Most recently, researchers have started to explore the role of heuristics on how people make moral decisions in the real world. In particular, under uncertain situations where the set of information is not always completely known, people are found to adopt heuristic strategies, which are different from processes of rational optimization. Some heuristics are believed more efficient with higher accuracy compared to complicated regression analysis, but others can lead to mistaken or even absurd moral judgments (Sunstein, 2005). The next section will focus on principles of moral heuristics from both perspectives above, how these principles have been applied to which situations, and the corresponding consequences of applying some principles.

Moral heuristics

Moral heuristics do not stand alone; rather, they stand within the realm of heuristics across decision-making domains. Heuristics—mental short-cuts, or rules of thumb, are often touted as efficient cognitive processes that ignore part of the information, with the goal of making

decisions more quickly, frugally, and/or accurately than more complex methods (Gigerenzer & Gaissmaier, 2011). As Tversky and Kahneman (1974) contended: "People rely on a limited number of heuristic principles which reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations". Heuristics can save effort and energy by ignoring part of the information, so they are usually considered "fast and frugal". Aligned with the view of Gigerenzer and Gaissmaier (2011) on heuristic strategies, moral heuristics reviewed in this section can be drawn upon consciously or unconsciously, unlike the dual-process models of reasoning that link heuristics to unconscious and error-prone processes (Kahneman & Frederick, 2005).

For non-moral decision-making under uncertainty, in particular, heuristics strategies have been found to describe people's actual decisions more accurately if heuristic strategies are appropriately selected (Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011; Hertwig & Herzog, 2009; Tversky & Kahneman, 1974). Two representative classes of heuristics will be summarized and distinguished below based on the following components: search rules (direction of searching in the search space), stopping rules (to what extend or when the search stops), and decision rules (how the final decision is determined) (Gigerenzer & Gaissmaier, 2011). Each component will implement a step of the procedure of heuristics.

Take-the-best heuristic selects the one that wins on the first attribute discriminating between alternatives. Take-the-best is a non-compensatory strategy, as it does not consider other attributes once a "winner" appears. The procedure for implementing this strategy is: (i) search through attributes and rank them in the order of validity; (ii) stop searching when the first attribute appears to discriminate between the alternatives; (iii) choose the alternative with positive or higher value on the first attribute.

Trade-off heuristic weights each alternative or attribute equally and thus is compensatory. For example, tallying, one strategy of the trade-off heuristics, compares the number of attributes or cues favoring one alternative over others. The procedure for tallying is: (i) search through attributes in any order; (ii) stop searching at n out of N attributes ($1 < n \leq N$). If the number of positive attributes is the same for both alternatives, continue to search for the (n + 1)-st attribute. If $n = N$ (no more attributes left in the search space), choose either alternative with the same probability (i.e., guess); (iii) choose the alternative which has more favorable attribute.

Moral heuristics are heuristic strategies people utilize in moral contexts, especially if the assumptions of rational analysis are not met. Similarly, moral heuristics do not require complete knowledge of the context, such as the likelihood of outcomes resulting from specific actions, nor do they require thorough computation based on available knowledge. Studies have shown that people adopt heuristics when making moral decisions from the perspectives of bounded and ecological rationality (Fleischhut, 2013; Gigerenzer, 2010; Sunstein, 2003; 2005). Bounded rationality refers to the fact that we have limited capacity for information processing given complicated inputs from the external world with uncertainty to some extent. Ecological rationality focuses on the interplay between our cognition and the environment we live in. These two perspectives are embodied in two questions researchers have been seeking to address: one is descriptive: Which heuristics do people use in which situations? The other is prescriptive: When should people rely on a given heuristic rather than a complex strategy to make better judgments? Most studies on moral heuristics have focused on the first question, which will be mainly focused in the remaining section. To address the first question, two principles of moral heuristics, some specific moral heuristics, reflections upon these heuristic strategies, and their application to biases in moral decisions will be discuss in the rest.

The two principles of moral heuristics respectively address two aspects of decisions under uncertainty: bounded rationality and ecological rationality. Our moral decisions are products of mind and the environment; by looking at either side alone, we cannot fully understand the real mechanisms underlying moral behavior. The first principle, "*less-can-be-more*" principle, mainly focuses on the computational processes in our mind that ignore part of the information thereby reduce our mental effort. It claims that more information or computation can decrease accuracy; therefore, to be more accurate, minds rely on simple heuristics over complex strategies (Gigerenzer, 2010; Gigerenzer & Brighton, 2009; Gigerenzer & Gaissmaier, 2011). Unlike the regression analysis that usually requires sufficient amounts of data and fits the data into models with multiple parameters, the less-can-be-more principle takes full advantage of available information and avoids problems like "overfitting" or "suffering from excess complexity" (Gigerenzer & Brighton, 2009). This principle can also be understood to evaluate moral situations based on restricted codifiable or uncodifiable moral principles. The more principles are engaged, the more likely it will generate conflicts between principles.

The second principle, "*Simon's scissors*" principle, takes into account the impact of the environment on moral decision-making processes. It claims that human rational behavior (and the rational behavior of all physical symbol systems) is shaped by a scissors whose two blades are the structure of task environments and the computational capabilities of the actor (Gigerenzer, 2010; H. A. Simon, 1990). Our behavior is a product of both mind and the environment. By analogy, moral behavior is a function of mind *and* the environment rather than mind alone—merely from moral reasoning based on moral principles or driven by moral intuitions. The principle of Simon's scissors can be applied to moral luck. Although moral behavior can be partially determined by the environment rather than solely controlled by individuals, we still make moral judgments about people's behavior that may not be in their control. Our reasoning on causal responsibility seems to integrate the influence of the external world on moral behavior. Four heuristics mainly focusing on the impact of the environment were suggested to guide people's moral behavior, as they are believed to foster social coherence (Fleischhut, 2013; Gigerenzer, 2010):

(i) *Imitating-your-peers*: do what the majority of your peers do.

(ii) *Equality heuristic (1/N)*: to distribute a resource, divide it equally.

(iii) *Tit-for-tat*: if you interact with another person and have the choice between being kind (cooperate) or nasty (defect), then: (a) be kind in the first round, thereafter (b) keep a memory of the most recent interaction, and (c) imitate your partner's last behavior (kind or nasty).

(iv) *Default heuristic*. If there is a default, do nothing about it. A representative example is organ donation programs across countries, in which default heuristic plays an important role.

These four moral satisficing principles seem not to level people's motivation and rationality the same as the environmental impact on moral decision. Indeed, the impact of the environment on moral decisions can hardly be exaggerated; however, we human, I would argue, are also motivated to reason to varying extent. For example, the principles "imitating-your-peers" and "default heuristic" do not allow much individual differences, though they provide us with easy and straightforward solutions, especially when we have to make difficult decisions. The "equality heuristic" appears reasonable following deontic rules, but it does not consider situations where individual contribution may vary. Absolute equal distribution might not be practical across situations. Also, under the principle of "tit-for-tat", first, judging whether the behavior of another person is kind or nasty is vague, as the judgment is merely based on the

consequences of the action the other person take. Also, the tit-for-tat strategy will lead immediate punishment after one single misconduct, which may generate unnecessary conflicts and does not seem forgiving like people are.

Besides, some heuristics have been found to produce moral mistakes in some contexts including in the domains of law and policy. These heuristic strategies work efficiently in most cases, though they can lead to biases based on incomplete evaluation by a selected set of moral principles. Note that the following cases are still under debate. I list some cases here to demonstrate the two questions asked at the beginning of this section are worthwhile examining, as we cannot apply moral heuristics to all situations in the real world. In cases of risk regulation (Sunstein, 2005), people exhibit discrepant moral judgments under the same heuristic strategies:

(i) Do not knowingly cause a human death. This heuristic strategy seems only to evaluate moral situations based on deontic rules that we ought not to harm others or treat others as a mere mean. Consider this example: a company knows that its product will kill ten people. It markets the product to its ten million customers with the knowledge. The cost of eliminating the risk would have been $100 million. People may judge that this company should be punished, as its analysis measures potential lives lost and monetary values. The problem is that it is not always unacceptable to knowingly cause death when the deaths are relatively few and unintended. For example, companies produce tobacco products knowing that many people will die due to tobacco consumption, and the government does not ban these products.

(ii) People should not be permitted to engage in moral wrongdoing for a fee. This heuristic indicates that moral transgressions should be tradeoff with monetary values. A problem of emissions trading arises, as some people believe that polluting the environment is the same as rape, theft, or battery, which should not be justified by licenses. The problem is that pollution might be different, as it comes as a byproduct of beneficial social activities.

(iii) People are especially averse to risks of death that come from products (like airbag) designed to promote safety. Similarly, Punish, do not reward, betrayals of trust. For example, people are not willing to buy airbags if they are informed that the rate of death caused by deployment of the airbag, though the rate is lower than the rate of death caused by car crash. The rational choice is actually to buy airbags. In the same vein, in a case where a babysitter neglects a child, people tend to punish the babysitter more harshly than someone who is not a beneficiary of trust, as they feel betrayed. The contrast between expectation and actual outcomes may trigger

people's outrage. This contrast might also be the reason that people judge someone blameworthy under the situation of moral luck, as people anticipate that good intent should lead to reasonably acceptable rather than negative outcomes. This heuristic is adopted when people's emotion plays an important role in moral judgments.

To sum up, moral heuristics account for both bounded rationality and ecological rationality in moral decisions, which are implemented through less-can-be-more and Simon's scissors principle, respectively. While moral heuristics can lead to controversial judgments under certain moral contexts, moral heuristics also serve as an efficient tool to facilitate moral decisions in the real-world, especially with difficult decisions or incomplete knowledge of the environment and/or social relations. Thus, moral heuristics may have both normative and descriptive functions whose adequacy is still in question. Also, when people should rely on a given heuristic rather than a complex strategy or self-motivation to make better judgments has yet to be answered.

The following two chapters extend moral decisions closer to our real life. The hypothetic, sacrificial, and deterministic moral dilemmas have been criticized for limiting the generalization of results and the actual underpinnings of mundane moral judgment on a daily-life basis (Bauman, McGraw, Bartels, & Warren, 2014; Bennis, Medin, & Bartels, 2010; Haidt & Joseph, 2004). Recently moral studies have been conducted in virtual reality to explore the potential discrepancy between moral judgments and moral actions (Hertwig & Erev, 2009; Hertwig, Barron, Weber, & Erev, 2008; Iliev et al., 2012; Navarrete, McDonald, Mott, & Asher, 2012; Patil et al., 2013; Skulmowski, Bunge, Kaspar, & Pipa, 2014). Moreover, in reality, many important moral decisions require the evaluation of choices involving both outcomes of variable magnitude and probability (Ruff & Fehr, 2014; Shenhav & Greene, 2010), which is similar to economic decisions. Moral situations we encounter in daily life often bear uncertainty to some extent, in which the consequences of actions and their probabilities are not clearly known, and have more than two alternatives we can choose from. However, the underlying mechanism of moral decisions under risk or with multi-alternatives has yet to be explored further. Since patterns of moral judgments are believed to be mediated by non-moral psychological representation (Cushman & Young, 2011), probabilistic and multi-alternative non-moral decision theory may shed light upon the process of real world moral decisions under risk and under uncertainty.

Chapter IV

From non-moral decision-making to moral decisions

Decision theories in this chapter are categorized based on the characteristics of decision situations and available information. "Characteristics of decision situations" here refer to the states of knowledge or forms of information under which decisions are made: certainty (the causal relations between acts and consequences are known and definitive), risk (the occurrence is certain or determined by specified random process), uncertainty or ignorance (either likelihood of outcomes occurrence or corresponding outcomes of acts are unknown), and number of options (e.g., binary or multi-alternative decisions). Most moral decision research has explored decisions under certainty with two options, as summarized above. In decisions under certainty, decision makers know the exact consequence of each choice and that the consequence will certainly occur. To illustrate, in the bystander trolley problem, one needs to choose to flip a switch to divert the train or not, and the consequence of the choice is certain and known- flipping the switch saves five persons for sure (Figure 3A).

However, moral decision situations we encounter in the real world often involve risk (Figure 3B) and/or offer multiple options. Figure 3B demonstrates two probabilistic variants of the classical trolley problems, where one action may have more than one consequence. For example, if a decision maker chooses to flip a switch, there could be a 10% chance that the switch does not work due to latency of the decision maker's response (i.e., it is too late to flip the switch) and a 90% chance that the switch would work and the trolley would be diverted onto the side track. In this case, the likelihoods (or risk) of consequences are known, but once the decision maker makes a decision, the actual consequence is determined by the external world. There is recent research on moral decisions under risk, such as manipulating the number of lives that can be saved with explicit probabilities of success (Shenhav & Greene, 2010). The comparison between options may rely on utilitarianism that saving more people is considered better than saving fewer. But further studies are necessary to reveal the mechanisms of moral decisions under risk when other codifiable and/or uncodifiable principles are involved besides utilitarian principles.

Furthermore, the likelihood of consequences or the structure of the situation may not be available at the time of decision (decisions under uncertainty). Suppose in a similar trolley

problem but with the likelihood of each consequence not explicitly known to decision makers, the decision makers may estimate the likelihood based on other information and past experience. Also, the consequence resulting from each choice may not be known in some contexts. For example, the decision maker may choose to push a large man onto the side track, but the consequence is not foreseeable. Multiple possibilities exist. The large man could fight against the decision maker, not fall on the track in time, fall on the track but not stop the trolley, and so on. Moral decisions under uncertainty as such are not uncommon in the real world. However, it's still an open question how our mind assesses the environment, estimates and integrates accessible information, and makes final choices.

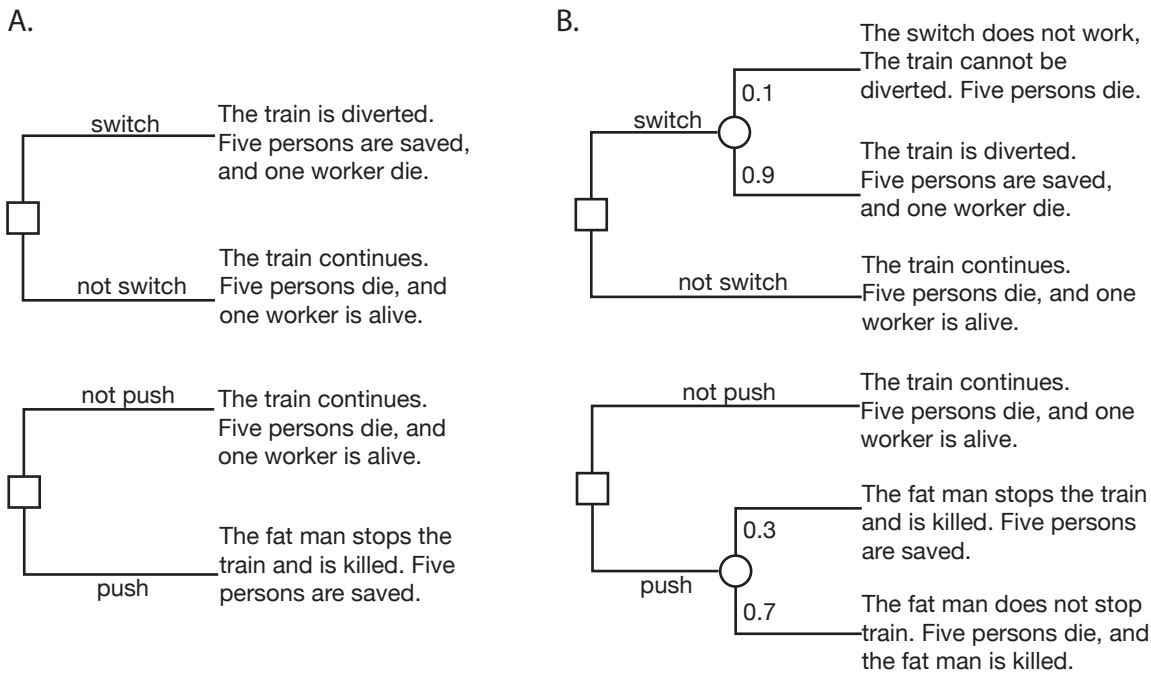

Figure 3. Decision tree for trolley problems. A. The classical trolley problems with certain consequences and two alternatives. B. An example of probabilistic trolley problems. The squares represent "choice points" at which a decision maker selects a course of action. The lines represent choices that lead to the consequences following the chosen actions. Circles represent consequences. In B numbers indicate the likelihood of each consequence. For example, in the upper panel of B, if the decision maker chooses to switch, there is a 10% chance that the switch does not work and a 90% chance that the switch works as expected.

Given that the moral decision under risk, under uncertainty, and with multi-alternative/multi-attribute has yet to be more elaborated on, this chapter will mainly focus on decisions theories under the same situations but in the non-moral domain. The aim is to provide the fundamentals and to inspire future studies on the underlying mechanisms of moral decisions in real life.

Decisions under risk

Decisions under risk have been often studied with choices whose values or outcomes and corresponding probabilities are definitive and known to decision makers. Unlike decisions under certainty where outcomes are fixed, decisions under risk are accompanied by 38robabilistic future states. Often gambling tasks are used to study risky choices. Gamble outcomes are determined jointly by the choice of the individual and the result of some specified random process (Coombs, Dawes, & Tversky, 1970).

Decision theory seeks to explain how decisions are made or ought to be made. It has normative and/or descriptive functions. Normative or rational decision theories attempt to specify optimal choices rather than actual choices, and to prescribe which decisions should be made to maximize the gain given the goal of decision makers and available information. Decision theories which were tended to provide normative[9] functions include the Expected Value Theory (EV), the Expected Utility Theory (EU) (Neumann & Morgenstern, 1947), and the Subjective Expected Utility model (SEU, see Section 4.2) (Savage, 1954). Different from EV which was primarily tended for normative purpose, EU and SEU are also important for descriptive applications, as both of them consider subjective transformation of objective values and/or probabilities. In contrast, descriptive decision theories attempt to explain how actual choices are made and the factors that affect those choices. Decision theories which were tended to provide descriptive functions include Prospect Theory (PT) (Kahneman & Tversky, 1979) and Cumulative Prospect Theory (CPT) (Tversky & Kahneman, 1992). PT and CPT can account for some choice behavior that violates predictions by EV or EU.

The main part of this section is devoted to theories of decision-making under risk: EV that directly uses objective values and probabilities to calculate optimal choices, EU that replaces objective values with subjective utilities, and PT and CPT that transform both objective values and probabilities to subjective utilities and subjective probabilities. SEU that estimates the likelihood of events will be discussed in Section 4.2, as SEU can be also applied to decisions under uncertainty. At the core of decision theories, two elements are of interest: values and

[9] I categorized these decision theories roughly based on their "normative" and "descriptive" functions, as the normative and/or descriptive functions of some are under debate. For instance, the expected value rule has been found inadequate on both normative and descriptive accounts. Also, normative and descriptive theories are deeply interrelated in most applications (Coombs et al., 1970).

probabilities of choices. These theories differ in whether or which subjective elements that are taken into account in the computation of optimization.

EV assumes we humans are rational and expected value maximizers and proposes the expected value rule – people compute the expected value of each option and select the one with the highest expected value. The expected value of a gamble or an alternative is the sum of the values of all its outcomes weighted by corresponding probabilities. More formally, the expected value of a gamble (G) with with n outcomes is:

$$EV(G) = \sum_{i=1}^{n} p_i x_i \tag{4.1}$$

where the values of outcomes are denoted by $x_1, x_2, \dots, x_n$ and each outcome can be obtained by probabilities $p_1, p_2, \dots, p_n$, respectively. Both values (x_i) and probabilities (p_i) are objective information. However, people do not always follow the expected value rule. In the first place, people are willing to pay for insurance to secure themselves against events with low probabilities. The insurance expense is often higher than the expected value of the unfavorable but low probability events, though people feel it is reasonable and rational to do so. Second, people are willing to accept gambles with negative expected values in a casino. From the two examples, the expected value rule seems inadequate to capture all characteristics of choice behavior. To capture the violation of EV, the latter two theories to be discussed instead take into account subjective perception of either objective values, or both values and probabilities, respectively.

Unlike EV, EU assumes that people perceive values in a nonlinear manner. As such, EU replaces the objective values in expected value rules with subjective utilities by a utility function $u(x_i)$. The curvature of the utility function reflects people's risk attitudes: risk averse individuals have concave utility function; risk-seeking individuals have convex utility function. The expected utility of a gamble (G) is computed as:

$$EU(G) = \sum_{i=1}^{n} p_i u(x_i). \tag{4.2}$$

To demonstrate the application of EU, suppose $u(x_i) = x_i^{\alpha}$ $(x_i \geq 0)$ for some positive α, which indicates the curvature of the utility function. The parameter α is often related to risk attitude of the decision maker, and also accounts for individual differences in utility transformation. If $0 < \alpha < 1$, the utility function is concave and the decision maker is considered risk-aversive. That is, the decision maker prefers the certain option (x_i) to any risky options with the expected value (x_i). If $\alpha > 1$, the utility function is convex and the decision maker is considered risk-seeking. If

$\alpha = 1$, the decision maker is considered risk-neutral, as the objective values are equal to the utilities.

While EU takes into account the subjectivity of value evaluation, there are some situations where people violate the prediction of EU. People often overestimate outcomes that are considered certain, relative to outcomes which are merely probable. This is called the certainty effect or sure-thing principle (Kahneman & Tversky, 1979), which is incompatible with EU. One counterexample of EU is the "Allais paradox" (Allais, 1990).

Situation 1. Choose between	Situation 2. Choose between
Gamble 1.	Gamble 1'.
0.5 million with probability 1;	0.5 million with probability 0.11,
Gamble 2.	nothing with probability 0.89;
2.5 million with probability 0.10,	Gamble 2'.
0.5 million with probability 0.89,	2.5 million with probability 0.10,
nothing with probability 0.01.	nothing with probability 0.90.

Most people choose Gamble 1 in the first situation but Gamble 2' in the second. To examine these two situations in detail, consider the following table:

		0.01	0.10	0.89
Situation 1	Gamble 1	0.5 million	0.5 million	0.5 million
	Gamble 2	0	2.5 million	0.5 million
Situation 2	Gamble 1'	0.5 million	0.5 million	0
	Gamble 2'	0	2.5 million	0

Both Gamble 1 (or 1') and 2 (or 2') have a chance of 89% to win 0.5 million in Situation 1 (or 0 in Situation 2) (see the third column in the above table). Under these two situations the other elements of two gambles remain the same. If a person prefers Gamble 1 under the first situation, based on EU he will also prefer Gamble 1' under the second situation, as the common component of both gambles would be discarded and not affect his choice. Based on common consequence principle that we should ignore the common consequences in options under consideration, if a person prefers Gamble 1 in Situation 1, he would prefer Gamble 1' in Situation 2. However, people often exhibit preference reversals. Under situation 1, people prefer Gamble 1, meaning

$$u(Gamble\ 1) > u(Gamble\ 2)$$

and thus

$$1\times u(0.5) > 0.10\times u(2.5) + 0.89\times u(0.5) + 0.01\times u(0).$$

By subtracting $0.89\times u(0.5)$ on both sides, we have

$$0.11\times u(0.5) > 0.10\times u(2.5) + 0.01\times u(0). \quad (4.3)$$

In the same vein, people prefer Gamble 2 under situation 2, meaning

$$u(Gamble\ 1') < u(Gamble\ 2')$$

and thus

$$0.11\times u(0.5) + 0.89\times u(0) < 0.10\times u(2.5) + 0.90\times u(0)$$

By subtracting $0.89 * u(0)$ on both sides, we have

$$0.11\times u(0.5) < 0.10\times u(2.5) + 0.01\times u(0). \quad (4.4)$$

Comparing 4.3 with 4.4, people's preference under the two situations is inconsistent with the basic tenets of EU. The reason of reversing preference may be that: Gamble 1 has certainty of winning 0.5 million without risk in Situation 1, so people prefer the sure option without taking risks; in contrast, both gambles in Situation 2 are under risk, so people compute the expected utilities of both and choose the higher one. Thus, EU is, in this case, not descriptively valid. The inconsistent decisions in Allais' paradox may be explained by anticipated regret (i.e., people would feel regret if they choose the other probable option but get nothing) (Tversky, Slovic, & Kahneman, 1990). Another explanation is that people compute the utilities weighted by nonlinear probabilities, which will be discussed next.

To explain the pattern of risky choices that the EU, PT was suggested by Kahneman and Tversky (1979) (Figure 4). Different from EU, PT assumes a discontinuous probability weighting function $\pi(p_i)$, identifies a reference point which is not necessarily zero, and allows distinct curvatures of value functions $u(x_i)$ in gain and loss domains. That is, PT maps values and probabilities into their subjective counterparts. The utility function of gamble G is:

$$U(G) = \sum_{i=1}^{n} \pi(p_i)u(x_i). \quad (4.5)$$

The probability weighting function of PT is discontinuous when probabilities are close to 0 or 1. It assumes high probabilities are often underestimated (Figure 4A, the solid line being above the dotted line as the probability p approaches 1), low probabilities overestimated (Figure 4A, the solid line below the dotted line as p approaches 0), and the crossover (locus of intersection of the solid and the dotted lines) is less than 50%. Hence PT has characteristics of subcertainty (i.e., weights for complementary events do not sum to 1) and subadditivity (i.e., for small probabilities

p_i and p_j, $\pi(p_i) + \pi(p_j) \geq \pi(p_i + p_j)$). The observed pattern in Allais' paradox seems reasonable by adding probability weighting function:

$$1 \times u(0.5) - \pi(0.89) \times u(0.5) > \pi(0.10) \times u(2.5) + \pi(0.01) \times u(0) > \pi(0.11) \times u(0.5)$$

and hence

$$1 > \pi(0.89) + \pi(0.11).$$

This result can be interpreted by the subcertainty of PT.

The value function of PT is concave in the gain domain, and convex in the loss domain with steeper slope than that in the gain domain. Also, a reference point accounts for the default state, or expectation basis, of decision makers, and thus the shift of value functions. If the decision maker expects to obtain a refund of $100, any refunds below $100 will be considered as a loss (which would be still considered as a gain in EU as long as the refund is greater than $0). In this case, the perception of gain and loss may vary across situations. While PT advances a nonlinear transformation of the probability scale in choices between risky options with a small number of outcomes, the theory does not provide computational solution to its fuzziness of weights near 0 and 1 or to the precise placement of the crossover.

CPT (Tversky & Kahneman, 1992), incorporates a cumulative functional into a gain-loss framing process of PT, and extends the applications of PT to risky choices with any number of outcomes. In CPT, the cumulative functional is a transformation on the cumulative probability distribution, instead on individual probabilities. The utility of an option in CPT is determined in the same manner as that of PT (Equation 4.3), but it provides a convenient mathematical representation of the probability weighting function. CPT assumes that there are two phases of decision-making processes: framing and evaluation phases. In the framing phase, decision makers construct the acts or choices contingent upon gains or losses. In the evaluation phase, each option is assessed depending on its domain from the framing phase. The value function[10] can be defined as:

$$v(x_i) = \begin{cases} x_i^{\alpha}, x_i \geq 0 \\ -\lambda(-x_i)^{\beta}, x_i < 0 \end{cases} \quad (4.6)$$

[10] Both value function and probability weighting function can have different variants. Here the two functions are provided as representative examples.

where α and β capture the shape of curvature of the value function in the gain and loss domain, respectively. The parameter α indicates the risk attitude of decision makers. The parameter λ indicates the loss aversion of decision makers.

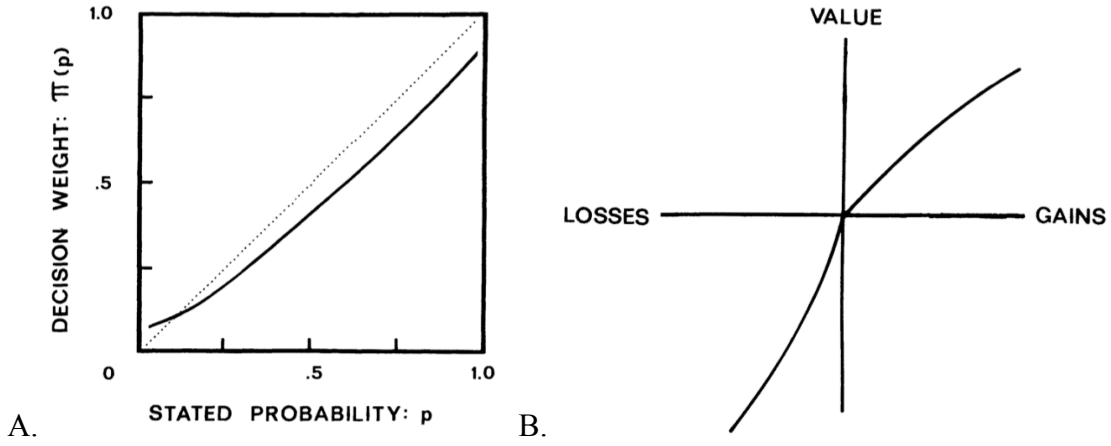

Figure 4. Probability weight function and value function of the Prospect theory. A. A hypothetical probability weighting function. B. A hypothetical value function. The intersect point of gain and loss curves is the reference point that divide the value into gains and losses. Adopted from Kahneman & Tversky (1979).

And the probability weighting function of CPT can be defined as (I use decision between two options as an example):

$$\pi(p_i) = \frac{p_i^c}{(p_i^c+(1-p_i)^c)^{1/c}} \tag{4.7}$$

where $c = \gamma$, if $x_i \geq 0$, and $c = \delta$, if $x_i < 0$. Note that the probability weighting function does not have to be restricted to two options and can be generalized to multiple options under consideration. The probability weighting function of CPT is continuous, whose curvature and crossover can be determined by the parameter c. Tversky and Kahneman (1992) found that people are generally risk-aversive as well as loss aversive, and overestimate low probabilities and underestimate high and moderate probabilities. Overestimating low probabilities contributes to the acceptance of lotteries and insurance. Underestimating high probabilities contributes to risk aversion in choices between certain and probable gains, and also risk seeking in choices between certain and probable losses.

The theories of decision under risk reviewed above have been tested not only in behavioral studies, but also neuroscience research in both moral and non-moral domains. Not surprisingly,

brain regions relevant to expected values of non-moral choices (mainly from risky gambles) are also associated with "expected moral value of actions", probable and certain outcomes of acts (Shenhav & Greene, 2010). The expected moral value of actions can be computed based on utilitarianism. The decisions in the experiment are to choose between two acts in a moral context: one default action that can save a few people with certainty, and the other action that can save a group of people (more than that the first action can save) but with known probabilities. Nevertheless, one of the big challenges of applying these theories in the moral domain is that consequence maximizing may not be considered optimal according to other moral principles, like deontology. That is, saving more lives cannot justify the harmful action toward an innocent. Thus, optimality in morality cannot be easily defined. Moreover, studies showing that people may perceive some moral beliefs more objectively than others raise the question how people integrate both magnitude and valence of moral transgression across situations (Goodwin & Darley, 2010). More studies on moral decisions under risk are necessary to explore this question and the relationship between different moral principles and rules. Also, under a single framework of moral principles (e.g., utilitarianism) how people construct the representations of moral problems has yet to be revealed. For example, whether people interpret the objective number of lives that can be saved and the probabilities in a linear or non-linear manner is still an open question.

Decisions under uncertainty

The assumption held by decisions under risk that the information of all alternatives, and their states and outcomes are available to decision makers to examine is, in most cases, unrealistic. First, situations where the likelihoods of future events are ambiguous are not uncommon. Second, the set of alternatives may be very large, or the structure of decision tree (Figure 3) may be very intricate, so that it may become practically unfeasible to have an exhaustive evaluation before decisions are made. Third, individual choice behavior is notorious for its inconsistency; people may make different decisions or judgments under seemingly identical conditions with the same set of alternatives. As such, decisions under uncertainty also concern decision makers' states of mind (unsure preference in the internal world). In the first part, literature on non-moral decisions under uncertainty will be discussed, from Ellsberg Paradox (Ellsberg, 1961), through evidence

illustrating that we take the uncertainty into account (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005; Huettel, Stowe, Gordon, Warner, & Platt, 2006), and to decision theories like SEU (Savage, 1954) and simplification rules for decisions under ignorance (Coombs et al., 1970). Decision theories on unsure preference will be included in the second part, such as constant utility model and random utility model.

Let us first consider the Ellsberg Paradox (Ellsberg, 1961). Imagine an urn known to contain 30 red balls and 60 black and yellow balls altogether, the latter in unknown proportion. One ball will be drawn at random from the urn. People are given two actions to choose from. The payouts in two situations are illustrated below. It is not hard to see that the outcomes of yellow balls are identical for both action 1 and 2 under both situations.

Situation 1: Action 1: "a bet on red" and Action 2: "a bet on black".

	30	60	
	Red	Black	Yellow
I	$100	$0	$0
II	$0	$100	$0

Situation 2: Action 1': "a bet on red or yellow" and Action 2': "a bet on black or yellow".

	30	60	
	Red	Black	Yellow
I	$100	$0	$100
II	$0	$100	$100

However, people often show discrepant preference under these two situations: a frequent observed pattern of response is action 1 is favored over action 2 under situation 1, while action 2' is favored over action 1' under situation 2. The pattern of choices may result from people's ambiguity aversion. In the Ellsberg Paradox, the proportion of black and yellow balls is unknown, which makes it difficult to assess Action 2 in Situation 1 and Action 1' in Situation 2. To bet on either yellow or black will require estimation of its proportion, which is unknown in this paradox. People seem to avoid ambiguity under these two situations by choosing the actions which do not require estimation of unknown probabilities. The Ellsberg Paradox illustrates the influence of ambiguity or uncertainty on people's choice behavior.

Neuroscience studies also have explored how ambiguity affects people's choice behavior and whether decisions under risk and under uncertainty evoke distinct neural circuits is still under

debate. Hsu et al. (2005) replicated similar ambiguity aversion in the Ellsberg Paradox and found that some brain regions were more active during ambiguous condition than that during risky condition and the ambiguity related regions are often believed to integrate emotional and cognitive input (Hsu et al., 2005). Later Huettel et al. (2006) suggested decisions under risk and under ambiguity involve two distinct mechanisms.

Given the complexity of decisions under uncertainty, relevant decision theories and simplification rules have been developed as discussed in the next section.

Decisions under ignorance[11]

There are circumstances where we have practically little or no confidence in the validity of our information, such as the chance of a terrorist attack. Expected utility maximization is in practice impossible. One approach, if the number of alternatives is not too large and the structure of decision trees is not too complex, is to estimate the likelihood of events with available information, or to assign unbiased probabilities to all alternatives when there is no reason to treat alternatives differently. The second approach is to circumvent likelihood consideration by adopting rules for decisions under ignorance. The third approach is to reduce the complexity of the problems, which contain incomplete information or have complicated structure, to manageable proportions by replacing the maximization principle by a weaker "satisficing" condition. These approaches can also be used for decisions with multiple alternatives and/or attributes (see more decision strategies in Section 4.3). The latter two approaches can simplify difficult decision problems without requiring comprehensive knowledge of probabilities of events. The section discusses these three approaches in turn.

According to EU, numerical probabilities are assumed to be known *a priori* to measure the utilities of outcomes. For situations where no a priori knowledge of numerical probabilities is available, SEU replaces objective probabilities with subjective probabilities which may be estimated by the likelihood or the corresponding events (Savage, 1954).

$$SEU(G) = \sum_{i=1}^{n} s(E_{x_i})u(x_i) \tag{4.8}$$

[11] Here ignorance is used to avoid confusion referring to that we do not have complete information of the external state, as uncertainty consists both ignorance and unsure preference.

where $s(E_{x_i})$ represents subjective estimation of the likelihood of Event x_i occurrence, which needs not to be specified in advance. SEU also allows individual differences in the simultaneous measurement of events (subjective probabilities) and outcomes (utilities). The common feature shared by both objective and subjective expected utility models is that the subjective value of an alternative, like a gamble, is a composite function consisting of desirability of the outcomes and the likelihood of events. While SEU can identify an optimal option based on the goal of decision makers and subjective estimation, it does not take into account the influence of confidence about these probabilities estimation on choices (Hsu et al., 2005). Also, it may not be feasible in some cases, where the structure of decision tree becomes complicated or the number of options are too large, it would cost much time and mental effort in computation and estimation. Therefore, some simplification rules for decision under ignorance have been introduced to facilitate the processes of decision-making (Coombs et al., 1970; Simon, 1957). Suppose the option set is $\{x_1, x_2, \dots, x_n\}$ and each option has m possible states or attributes $\{s_1, s_2, \dots, s_m\}$. The value or utility of the outcomes resulting from choosing x_i given state s_j is denoted by v_{ij}.

1. The maximin criterion: choose the alternative whose lowest value is the highest among the lowest values of all alternatives under consideration.
2. The maximax criterion: choose the alternative whose highest value is the highest among the highest values of all alternatives under consideration.
3. The pessimism-optimism criterion: assume the maximal and minimal values of alternative x_i are v_i' and $v_i''$and r $(0 \le r \le 1)$ a pessimism-optimism index. Choose the alternative x_i with the highest value of $rv_i' + (1-r)v_i''$.
4. Principle of insufficient reason: assign equal subjective probabilities (equally probable) to all states and then calculate the values of alternative x_i as $\frac{1}{m}\sum_{j=1}^{m} v_{ij}$. Choose the alternative x_i with the highest value.
5. The minimax regret criterion: assume the maximal value that can be obtained under state s_j as $\dot{v}_j$. The original v_{ij} is transformed to $v_{ij}' = \dot{v}_j - v_{ij}$, called regret value. Choose the alternative x_i whose maximal regret value is the lowest.
6. The satisficing principle is to choose the first alternative that is considered "satisfactory" with respect to all relevant attributes. That is, set satisfactory levels of all possible states,

and compare each attribute of x_i with corresponding satisfactory levels. Choose the first x_i that satisfies all preset levels.

Decisions under ignorance in the realm of morality are often seen in daily lives, so it would be interesting to examine whether and how people make moral decisions under ignorance. First, from the perspective of SEU, the estimation of outcome likelihood may be largely influenced by both acts and outcomes. For instance, people often refuse to push a large man onto the track may due to underestimating the success rate of the action, compared to flipping a switch which seems more likely to obtain a desirable outcome (save five persons). Second, people may also utilize simplification rules consciously or unconsciously in moral contexts and different rules may lead to different choices. More specifically, the maximin criterion may help people to avoid the worst outcomes. The maximax criterion may help people identify the best possible outcomes. The pessimism-optimism criterion mediates both the best and worst outcomes of alternatives by weighting to these two outcomes separately and guides people to find a compromise alternative. The principle of insufficient reason assumes the likelihood of outcomes resulting from an act is equal, e.g., the likelihood of two outcomes caused by push would be assumed 50% in Figure 3. The minimax regret criterion is similar to the maximin criterion, but the worst is defined by the maximal regret. Last, the satisficing principle requires people to decide satisfactory levels of all states or attributes of an alternative and then to choose the first one that meet these levels. Similarly, people may set thresholds of moral acceptance for each attribute, like assessment of outcome (e.g., how many persons to save) and action (e.g., whether the action directly causes harm), and then choose the first alternative that satisfies all thresholds. Note that the values in the non-moral decisions may differ from those of moral decisions, as the assessment of each attribute of each alternative can vary depending on which moral principles are applied. Also, which moral principles are selected in the assessment of alternatives relies not only on specific moral contexts but also on individual decision makers.

Decisions with unsure preference

Discrepancy has been often seen in individual choice behavior under seemingly identical conditions. The discrepant choices may reflect uncontrolled momentary fluctuations, like attention shift, or an inherently probabilistic choice mechanism (Coombs et al., 1970). To

interpret the fluctuations in psychological estimation of alternatives, Thurstone (1927) introduced the law of comparative judgment. This model assumes that alternatives are represented as distributions, or random variables along a common underlying dimension, and the probability of selecting one alternative over another is the probability that the first randomly drawn variable exceeds the second (Thurstone, 1927).

Decisions with unsure preference are often addressed by probabilistic theories of choice, which, in general, have two types of models, constant utility models, and random utility models. In constant utility models, the utility value of each alternative is fixed or constant and the probability of choosing one alternative over another is a function of their utility value differences. Different probability choice theories have been used (summarized in Scott, 2006). Constant utility models assume that choice behavior is determined by a random process. One example is Luce's constant utility model, where the probability of choosing x_i among all available alternatives $X = \{x_1, x_2, \dots, x_n\}$ is defined as:

$$p(x_i; X) = \frac{u(x_i)}{\sum_{j=1}^{n} u(x_j)} \tag{4.9}$$

In other words, the probability of choosing x_i among all alternatives depends on the ratio of its utility to the sum of utilities of all options.

In random utility models, the alternative with the highest utility is always chosen, but the utilities are not constant but random variables. As such, the utility of an alternative can vary from time to time, though the choice mechanism is deterministic. A simple assumption of the randomness of alternative values is that these values are selected from normal distributions (Thurstone, 1927), though more complex mechanisms are possible. Coombs' random utility model is an example of random utility models. For example, a decision is to select of a comfortable temperature in a room (Coombs et al., 1970). The ideal temperature level of every individual may fluctuate from moment to moment because of internal factors. Also, the sensation of a given temperature can vary over time. Suppose U_{x_i} and U_{x_j} represent the random variables of the sensation of two given temperature conditions x_i and x_j, and I represents the random variable of an ideal temperature. To choose one temperature between two conditions, the probability of selecting x_i over x_j is equivalent to the probability that the distance between x_i and the ideal is less than the distance between x_j and the ideal:

$$p(x_i, x_j) = p(\left|U_{x_i} - I\right| \leq \left|U_{x_j} - I\right|), \tag{4.10}$$

where $\left|U_{x_i} - I\right|$ indicates the distance between the given temperature and the ideal. On the other hand, in cases with two options, Equation 4.8 can be written for the binary preference probabilities (Thurstone, 1927):

$$p(x_i, x_j) = p(\left|U_{x_i}\right| \geq \left|U_{x_j}\right|). \tag{4.11}$$

These two types of models may be used to account for some flexibilities observed in moral decisions. People may have incomplete evaluation of actions under certain conditions (e.g., time pressure), like by ignoring some aspects of actions. Incomplete evaluation contributes to smaller difference in evaluated values between utilitarian and deontological choices (Suter & Hertwig, 2011). This may be explained by constant utility models, that the incomplete evaluation may lead to smaller discrepancy between probabilities of utilitarian and deontological judgments, as the probabilities of choosing two alternatives become closer to each other according to Equation 4.7. An alternative explanation is based on random utility models, that the variability of the distribution of evaluated values might become greater under certain conditions. Thus, the difference between probabilities of utilitarian and deontological judgments would decrease according to Equation 4.9. Understanding whether the observed inconsistency in moral decisions is related to internal unsure preference, or probabilistic choice mechanisms, is necessary to reveal the nature of moral decision-making processes.

Multi-alternative/multi-attribute decisions

Not all moral problems are binary; instead moral problems we are facing in the real world often have multiple solutions. However, studies on the mechanisms under moral decisions with multiple alternatives and/or attributes are scanty. Theories in the domain of general decision-making may shed light upon the potential mechanism of moral judgments with multiple alternatives. In this section I assemble theories and models on multi-alternative/multi-attribute non-moral decision-making on preferential choices where the outcomes are deterministic (Roe, Busemeyer, & Townsend, 2001; Usher & McClelland, 2001; 2004; Trueblood, Brown, & Heathcote, 2014) or probabilistic (Farmer, Warren, El-Deredy, & Howes, 2016; Wedell, 1991), followed by their applications on context effects.

When there are multiple options to choose, the evaluation of the options may depend on each other. This relation can result in choice behavior violating some fundamental assumptions of

decision theories. Tversky developed the Elimination-by-aspects (EBA) theory to remedy the inadequacy of classical probability theories of multi-alternative decisions (1972). EBA is a descriptive theory that explains our violation of utility theory under certain conditions when we make difficult decisions. We keep shifting back and forth favoring one option and another at next moment, until final decisions are made. In EBA, the subject's attention is allocated on the aspects that characterize the alternatives. If an aspect is only shared by a subset of alternatives, then all of the alternatives that do not possess this aspect are "eliminated" from the choice process. An "elimination" process of selecting an option is the outcome of a sequential selection of aspects. Such a sequence, as a "state of mind", could be understood as a sequence of attention shifts.

Inspired by the attention allocation process in EBA, several models have been developed to describe the mechanisms underlying multi-alternative/multi-attribute decision-making. The models to be discussed include Leaky Competing Accumulator (LCA) model (Usher & McClelland, 2001; 2004), Multialternative Decision Field Theory (MDFT) (Roe et al., 2001), and Multiattribute Linear Ballistic Accumulator (MLBA) model (Trueblood et al., 2014). These models attempt to describe people's decision processes in a *dynamic* manner. All three models consider choice behavior as gradual accumulation of evidence, within which each alternative accumulates its own evidence. A selected option is either an alternative whose evidence accumulation reaches a threshold first (no time constraint) or the one with the highest level of evidence at a constrained moment. Both the LCA and MDFT assume that evidence accumulation is leaky. The discussion of each model below includes their distinctive structures, especially attention allocation mechanisms, advantages and limitations, and interpretations of three main context effects. The three models discussed below assume that there are three alternatives X, Y, and Z, each characterized by two attributes P and Q.

Leaky Competing Accumulator (LCA) Model The LCA model is a dynamic, evidence accumulation model, which proposes that alternatives accumulate their evidence with leakage, self-excitation, and lateral inhibition against each other (Usher & McClelland, 2001; 2004). Decisions are made with two stages. First, the input preprocessing stage computes an input value of each alternative. Second, the leaky-integration stage propagates the input values to obtain activation values of the alternatives. The key components to be discussed include: loss-aversion

involved in the formulation of input values, leakage of integration, lateral inhibition, and attention shifts.

At the preprocessing stage, each alternative i $(i = X, Y, Z)$ is assigned an input value $I_i(t)$, varying with time. For example, the input value of X is

$$I_X(t) = F(d_{XY}) + F(d_{XZ}) + I_0 \tag{4.12}$$

where I_0 is a positive constant representing the "self-promoting" value that is uniformly pre-fixed for all alternatives; d_{ij} is the advantage or disadvantage differential of the option i relative to option j; and F is a nonlinear asymmetric advantage function, which captures loss aversion[12], defined as:

$$F(x) = \begin{cases} \log(1 + x), x > 0 \\ -\{\log(1 + |x|) + [\log(1 + |x|)]^2\}, x < 0 \end{cases} \tag{4.13}$$

A key property of F is that it has greater slope for negative values than that for positive values, which results in advantage for similar options and penalizes dissimilar option pairs.

At the leaky-integration stage, each option's input value is implemented into its activation value function $A_i(t)$, which describes its evidence accumulation from time t to $t + 1$. For alternative X, it is given by

$$A_X(t + 1) = \lambda A_X(t) + (1 - \lambda)\{I_X(t) - \beta[A_Y(t) + A_Z(t)] + \xi_X \cdot t\}, \tag{4.14}$$

where λ is a decay constant indicating the leakage of the integration, which reflects the memory loss of the alternative from the previous moment; □ is the global inhibition parameter; and ξ is a noise term from a normal distribution. At each moment within the integration stage, the subject's attention is allocated to one of the attributes (P or Q). The attention weights, w_P and w_Q,take values at either 0 or 1, i.e., an "all-or-none" pattern ($w_P(t) = 1$ and $w_Q(t) = 0$ or vice versa). The attention shift with time is determined by a pre-fixed probability $p(w_P)$ and $p(w_Q) = 1 - p(w_P)$. At time t, if the attribute P receives attention, then the advantage differential between two alternatives is the difference of their values along P.

The nonlinear advantage function can potentially limit its application in simulation due to computation intensity in the externally controlled paradigm (decisions need to be made at a fixed time) (Trueblood et al., 2014). Also, in an inference experiment and perceptual object decision tasks there is no notion of gains and losses on which loss aversion can operate (Trueblood et al.,

[12] This is consistent with PT and CPT (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992).

2014). Hence it might be limited to apply the LCA model as it attributes the similarity effect to loss aversion.

Multialternative Decision Field Theory (MDFT) The MDFT assumes that the preference for an alternative evolves over time through a series of evaluations and comparison of alternatives (Roe et al., 2001; Trueblood et al., 2014). The MDFT and the LCA differ in competitive interactions (lateral inhibition) among alternatives and loss aversion. The MDFT interprets the strength of lateral inhibition among options as a decreasing function of the "distances" between them. The MDFT first assigns a valance value to each alternative, which varies with time and is constructed from three components: (1) subjective values (personal assessments of the options), (2) stochastic attention weights (decision maker assigns weight to each attribute, varying with time), and (3) a comparison mechanism. At the integration stage, the MDFT applies a "preference state" function to each alternative. A decision is determined at the alternative whose preference state reaches a threshold. The following discussion focuses on the formulation of the valence values, the attention allocation distribution, and the "interconnection" between alternatives.

First, each option is evaluated with respect to the attributes P and Q, which can be conveniently represented by a matrix of subjective values:

$$M = \begin{bmatrix} m_{PX} & m_{QX} \\ m_{PY} & m_{QY} \\ m_{PZ} & m_{QZ} \end{bmatrix}. \tag{4.15}$$

The MDFT assumes that the subjective values of alternatives with respect to each attribute are based on the experience or knowledge of decision makers.

Second, the subjective values are weighted to form the weighted values of alternatives through the all-or-none attention allocation mechanism. The attention weights are represented by a vector:

$$W(t) = \begin{bmatrix} w_P(t) \\ w_Q(t) \end{bmatrix} \tag{4.16}$$

with which the weighted values of the options are computed by $MW(t)$.

Third, the valence value of each option is determined by the difference between the weighted value of the option and the average of the weighted values of all other options. Equivalently, this comparison mechanism can be summarized by a "contrast matrix", which in this case is:

$$C = \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{bmatrix}. \tag{4.17}$$

More generally, with n options it is an $n \times n$ symmetric matrix with value 1 along the diagonal and $-1/(n-1)$ off the diagonal. Last, the valence values of all options can be represented by the following vector $V(t)$[13]:

$$V(t+1) = CMW(t+1) = \begin{bmatrix} v_X(t+1) \\ v_Y(t+1) \\ v_Z(t+1) \end{bmatrix}. \tag{4.18}$$

At the integration stage, the preference states are determined by the valance values and competition among alternatives. For example, the preference of X is:

$$P_X(t+1) = S_{XX}P_X(t) + S_{YX}P_Y(t) + S_{ZX}P_Z(t) + v_X(t+1). \tag{4.19}$$

The weights S_{XX}, S_{YX}, S_{ZX} varying within 0 to 1 represent the "feedbacks" of the respective alternatives to X. The "self-connection" S_{XX} reflects the person's memory of the previous preference state of X. $S_{XX} = 1$ indicates that she has "perfect memory" and $S_{XX} = 0$ indicates that she has "no memory". The "interconnection" S_{YX} and S_{ZX}, typically negative, indicate lateral inhibitions. These *local* weights play a similar role as the *global* inhibition parameter β in LCA. The strength of S_{YX} is governed by the "distance" $Dist_{XY}$ between X and Y in the multi-attribute space[14], which satisfy two properties: (1) symmetry for a pair of options, i.e., $S_{XY} = S_{YX}$, and (2) decrease with respect to distance.

Without the nonlinear advantage function in the LCA, it is easier to fit the MDFT with experimental data in the externally controlled paradigm. Nevertheless, the MDFT has its own limitations. Mathematically, it is possible that the preference states are unbounded, which results in long stopping time; this issue can be resolved by adjusting parameters.

[13] If multiple attributes are involved, it would be practically desirable to group the attributes into two classes: a relatively small group of primary attributes; and a group of the remaining less relevant attributes. Then the valance value of an alternative is the sum of the primary component with a stochastic error component due to the less relevant attributes.

[14] The distance can be defined as $Dist_{XY} = \sqrt{(\Delta I)^2 + \beta \cdot (\Delta D)^2}$, where ΔI and ΔD are the differences of X and Y along the indifference and the dominance dimensions, respectively, and β is the dominance dimension weight.

Multiattribute Linear Ballistic Accumulator Model (MLBA) The MLBA is also a dynamic evidence accumulation model, but several features distinguish it from the other models. First, it explicitly incorporates a mapping from objective to subjective values. Second, it has a different attention allocation mechanism; it assigns to each pair of alternatives an attention weight along each attribute, measuring the similarity of the alternatives. The MLBA consists of pre-diffusion and diffusion[15] stages.

At the pre-diffusion stage, the subjective value m_{Pi} of alternative i with respect to attribute P is transformed from its objective value in a "curved" manner. First, a pair of alternatives X and Y are plotted in the two-dimensional attribute space at (P_X, Q_X) and (P_Y, Q_Y), respectively. The line connecting them represents all the alternatives "indifferent" from them. Second, the subjective value of any alternative Z on the indifferent line is the intersection of the ray from the origin of the plane through (P_Z, Q_Z) with the curve

$$\left(\frac{x}{a}\right)^m + \left(\frac{y}{b}\right)^m = 1 \tag{4.20}$$

where a and b are the P- and Q- intercepts of the indifferent line, and m is the curvature constant, uniformly chosen for all pairs of alternatives. If $m = 1$, then the alternatives' subjective values are the same as their objective values. If $m > 1$, then the subjective value of an intermediate alternative will be greater than its objective value, and is more preferred than extreme ones.

The MLBA directly compares all pairs of alternatives. For two alternatives X and Y, the valuation function v_{XY} of Y relative to X (here the comparison direction is important, as v_{XY} is not assumed to be equal to v_{YX}) is given by

$$v_{XY} = w_{PXY} \cdot (m_{PX} - m_{PY}) + w_{QXY} \cdot (m_{\square X} - m_{QY}), \tag{4.21}$$

where the coefficients w_{PXY} and w_{QXY} are attention weights. The attention weights reflect the similarities among the options. Fixing an attribute, the attention weight of a pair of options increases with their similarity. They are determined by

$$w_{PXY} = \exp(-\lambda\, |m_{PX} - m_{PY}|), \tag{4.22}$$
$$w_{QXY} = \exp(-\lambda\, \beta |m_{QX} - m_{QY}|).$$

[15] See Section 4.1 for a brief introduction of the diffusion process.

Where $\beta > 0$ indicates a bias toward attribute Q when $\beta > 1$ and a bias toward attribute P when $\beta < 1$. The λ represents the asymmetry in the mutual comparison. If $m_{PX} - m_{PY} > 0$, λ is set at λ_1, otherwise it is set to be λ_2, both are greater than or equal to zero.

At the diffusion stage, the mean drift rate of each option is determined by the sum of the comparison functions. The mean drift rate of alternative X is

$$d_X = v_{XY} + v_{XZ} + I_0, \tag{4.23}$$

where I_0 is a non-negative constant as a baseline input that plays a similar role as the I_0 in the LCA model. The constant I_0 ensures that at least one of the outcome mean drift rates is positive so that the model can be terminated within a finite period of time.

The MLBA overcomes some limitations of the LCA and MDFT. Regarding LCA, the loss-aversion assumption obstructs it from being further generalized to non-hedonic paradigms such as perceptual decisions. Regarding MDFT, it is rather challenging to explain the compromise effect, and it has some stability issues. In contrast, the MLBA is computationally simpler than the other two models; it is capable to explain a wider range of experimental paradigms; and it can be used to analyze the influence of deliberation time on choice preferences (Trueblood et al., 2014).

As deterministic multi-alternative/multi-attribute decision making extensively studied, there is also a line of research on paradigms in which alternatives possess probabilities (Wedell, 1991; Farmer et al., 2016). In particular, among the context effects, preference reversal due to attraction effect has been observed in multi-alternative probabilistic decision-making. In their paradigms, each stimulus consists of a pair: a monetary value and a probability. Then the product of the two components of each stimulus is its "expected value". When a decoy that is dominated by one of the two targets is added to the choice set, people could be led to choose the target dominating the decoy. In view of such a result, Wedell argued that people might have adopted a heuristic strategy in making decisions violating value-maximizing, called the dominance-valuing model. In this model, the decoy is perceived in a dominance relationship, which "directly increase the global attractiveness of the target" (Wedell, 1991). It is necessary to assume that the dominance relationship needs to be asymmetrical—if the decoy were dominated by both targets, the effect should not be salient. Moreover, the chosen target could be sub-optimal, with a lower expected value than the unchosen competitor (Farmer et al., 2016). Farmer et al. (2016) argued that the violation of maximizing expected values does not imply humans are irrational, as the

size of the attraction effect decreased as the expected value difference of the unchosen competitor over the chosen target increased (both absolutely and proportionally). Instead, people might still be aware of the ordinal relationships among the expected values of the alternatives. For example, if there are three alternatives, then there are three possibilities of the orders of their expected values, in two of which the target has higher expected value than the unchosen competitor, and this explains the attraction effect.

Similarity, compromise, and attraction effects These four theories summarized above can provide descriptive explanation for the three main context effects discussed in Chapter 2 (Figure 1A). It is worth noting that all the three models make similar predictions on the attraction and compromise effects, which becomes more salient as deliberation time increases. But they deviate on their predictions on the similarity effect. Suppose two cars are under consideration: car X is economic but of lower quality, and car Y is more expensive but of high quality. The preference of car X and Y may be altered when a third car is added. Similarity effect has been observed when adding car S that is similar to X in both price and quality. The EBA measures the similarities of options in terms of their attributes. The probability of choosing the dissimilar car Y is the highest, as car Y has more distinct attributes. The LCA explains the similarity effect using the leakage of integration. Even though loss aversion initially penalizes the dissimilar alternative Y, but alternative S shares the attention with the similar car X, contributing to an advantage for Y. Over time the leakage of the evidence integration will further amplify the advantage of Y. The MDFT compares the cars by their distances along dominance and indifference dimensions. In the indifference/dominance space, either X or S is more distant with Y than their mutual distance. This generates greater inhibition between X and S, and Y will be favored. The MLBA captures the similarity effect by predicting the updates on the mean drift rates of the options. If the decay constants satisfy $\lambda_1 < \lambda_2$, then the positive evidence will receive more weight than the negative evidence in the value functions. This may happen when people weight supportive evidence more than unsupportive evidence. Since the option Y is more dissimilar with S than X, the value v_{YS} will be greater than v_{XS}, and the mean drift rate of Y will be greater than that of X.

The compromise effect was observed if both of a third car C comes into consideration, which serves as a compromise between car X and car Y (Figure 1A). The LCA interprets the compromise effect as a consequence of loss-aversion. As car C is less distant to either X or Y

than X with Y, it receives the least penalty by the loss-aversion asymmetric function than both X and Y, thereby it is preferred. Also, the applicability of the MDFT to explaining the compromise effect may be limited (Usher & McClelland, 2004). The MLBA explains the compromise effect by not only the attention weights, but also more directly by the curvature parameter m. When people place a higher subjective value on the midrange car C, i.e., $m > 1$, then the compromised car C is preferred than either the extreme cars.

The attraction effect favors the car X when the third car A is dominated by X—of similar quality but cheaper. The LCA explains the attraction effect using the loss-aversion function. While the car X has Y as the only distant alternative, Y has both X and A as its distant alternatives, hence more significantly penalized by the asymmetric loss-aversion function. The MDFT views the attraction effect via a "bolstering role" played by the dominated car A to X in two steps. First, car A first suffers a negative preference state P_A due to its inferiority to X. Second, P_A feeds more significantly into a negative inhibition to P_X than P_Y to P_X. From the viewpoint of the MLBA, the attention weights w_{PXA} and w_{QXA} would be larger than w_{PXY} and w_{QXY} respectively due to the proximity of the alternatives X and A, and this is because the cars X and A are more difficult to discriminate than Y and A.

Compared to the literature on multi-alternative and multi-attribute decision-making in the non-moral domain, research on how people make moral decisions when they are faced with multiple options featured by multiple attributes remains at a preliminary stage. It is not uncommon that people make moral decisions with more than two options, and sometimes tend to think of alternatives during deliberation. Therefore, it would be important to extend current moral judgment theories to a multi-alternative and multi-attribute setting. The models discussed in this section might be as well applied in the realm of moral decisions.

CHAPTER V

Bridging moral and non-moral decision-making domains

Moral decisions in real life are more complicated than the hypothetical moral dilemmas extensively studied in the laboratory. To better understand moral decisions under risk, under uncertainty, and with multiple alternatives and/or attributes would be the first steps to reveal the true underpinnings of real-world moral decision-making. This chapter is intended to propose some experimental ideas and examples under the same framework of decision theories in non-moral domain.

Moral decisions from two deterministic alternative dilemmas to quasi-real-world problems

Moral decisions under risk

The impact of risk on decisions is prevalent, but moral decisions under risk have not been fully studied. Risk consideration can be reflected in forms like counterfactual thoughts (Byrne, 2016) and evaluation of "expected moral values" (Shenhav & Greene, 2010). Evidence supporting that people have counterfactual thoughts has been reported (Byrne, 2016). For example, in the bystander trolley dilemma, people tend to believe that the single person on the sidetrack would escape before the trolley hit him. Similarly, in the footbridge trolley problem, people may think that it is impossible for that large man to stop the trolley, in which harming the innocent would lead to even worse consequences. In both cases people still consider the likelihoods of certain events and other possible alternatives, even when such information is explicitly excluded in the classical trolley problems. Also, a recent study explored how people make moral decision by presenting dilemmas with known probabilities and found that the complex life-and-death moral decisions were related to more basic neural circuitry of decision-making (Shenhav & Greene, 2010). To further examine whether moral decisions are determined by a unique mechanism or by similar mechanisms of non-moral decisions under risk, probabilistic moral dilemmas (Figure 3B) and moral version of classical phenomena such as the Allais' paradox (Figure 5) would be worthwhile studying. First, we can test the existent decision theories in probabilistic moral situations and explore whether similar biases are also present in

moral decisions. The moral version of Allais' paradox is not exactly the same as the original one (the outcomes of gambles are non-negative), as moral situations often involve harm and sacrifice, rather than gains. People might treat moral situations differently, which may be explained by the certainty principle and loss aversion.

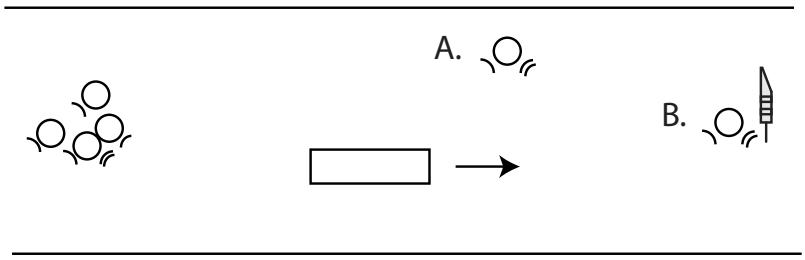

Figure 5. Moral version of Allais' Paradox. You are driving a small boat toward east in a deep river with intense, powerful, and somewhat violent rapids and notice two drowning men who are in front of you A and B. Person A is closer to you than person B. You receive a distress signal informing you that a small boat has capsized far away in the opposite direction, and all people aboard are now drowning. You know the only other rescue boar in the area is much further to the west, so would be unable to reach either A or B. But there is a chance the rescue boat will reach the group drowning to the west. The rescue boat can only be notified by handheld flares, which is held by Person B. Person B is also drowning, so it is impossible that he can signal the rescue boat to save the group by himself. **Situation 1**: You need to decide between two solutions: I. if you decide to save person A who is nearby, you will be able to save him in time, but it is impossible to save person B in time and the group of people will certainly die without launching person B's flares. II. If you decide to save person B and go full speed, it will be too late to drive back to save person A. You have an 10% chance to save person B and successfully signal the rescue boat to save the group of people in time; a 89% chance to save person B but it is too late to signal the rescue boat, so you will not be able save the group of people. However, as the rapids in the branch are hard to predict, there is a 1% chance that you will not reach person B in time, so you will not save anyone. **Situation 2**: You need to decide between two solutions: I'. if you decide to save person A who is nearby, you have a 11% chance to save him in time, as the rapids in the river are unpredictable. But you will not be able to save person B in time and the group of people will certainly die without launching person B's flares. II'. If you decide to save person B and go full speed, you have a 10% chance to reach him in time, as the rapids in the river are unpredictable. Once you save person B, you will be able to signal the rescue boat and the group of people will be certainly saved. This example is modified from Shenhav & Greene (2010).

Second, whether people perceive risk in the same way as they do during economic decision-making needs further exploration. For example, a probabilistic variant of the trolley problem may be one step toward a more naturalistic paradigm that reconciles with the real mindset of participants (Figure 3B). Also, studying how people transform objective value and probabilities

to subjective values and probabilities in moral situations can also help specify the relationship between moral and non-moral decision-making under risk. Number of lives and success rates of actions of each alternative, and "expected moral values" may be manipulated in a wider range to reveal the pattern of value and probability perception in moral decisions. More generally, the validity of decision theories under risk, such as the expected utility theory, the PT, and the CPT, in moral context has yet to be tested.

Last, utilitarianism is not the only moral principle governing moral decisions. That is, consequence optimization cannot be applied to all situations, such as in moral luck cases, as the outcomes are not intended by the person. In those cases, deontology and causal theory may influence more the way people make satisfactory moral decisions. In other words, to evaluate moral choices under risk based on the principles under deontological or causal theory framework may be more efficient and useful to determine moral accountabilities and blameworthiness.

Moral decisions under uncertainty

People prefer the known over the unknown in economic decision-making (Huettel et al., 2006). Whether this preference still holds in moral decisions is unknown, especially in moral dilemmas which often involve sacrifice and harm toward a few. The influence of uncertainty or ambiguity on moral decisions may differ from that on non-moral decisions, as people may consider lives that can be saved in a different way compared with monetary gains or losses. Moreover, the uncertainty in moral situations can be derived from limited information of the external state or unsure preference of the internal state sources of uncertainty (Coombs et al., 1970; Fleischhut, 2013).

An intriguing problem, known as the mineshaft problem (Parfit, 2011) may be a good candidate through which the processes of moral decisions under uncertainty with multiple alternatives can be explored. The mineshaft paradigm asking people to balance among three options is another perfect blend of rational reasoning and risky decision-making. To be specific, by choosing Gate 1 or Gate 2 the decision maker faces uncertainty about where these miners are trapped, while choosing Gate 3 guarantees that 90 miners will be rescued and 10 miners will be killed.

Mine Shafts[16]: A hundred miners are trapped underground, with flood waters rising. We are rescuers on the surface who are trying to save these people. We know that all of them are in one of two mine shafts, but we don't know which. There are three floodgates that we could close by remote control. The results would be these:

		The miners are in	
		Shaft A	Shaft B
	Gate 1	We save 100 lives	We save no lives
We choose:	Gate 2	We save no lives	We save 100 lives
	Gate 3	We save 90 lives	We save 90 lives

From the view of deontological principles, people may want to choose between Gate 1 and 2 but not Gate 3, as we ought to treat all humans equally and we should not sacrifice any person to save more. It is difficult to decide between Gate 1 and 2, since in which shaft these miners are trapped, is not available. From the view of utilitarianism, the optimal choice is to choose Gate 3 that can save out 90 miners for sure even at the cost of the other 10 miners. This problem posts new challenges to moral psychologists. First, few studies have explored moral decisions under uncertainty so far, though many moral situations we are faced with in real life contain uncertainty to some extent. This is the uncertainty from the external world. Second, while we may convert the uncertainty to probabilities with more information available to us, how the constructed representation of this moral situation influence moral decisions and whether the "expected moral value" is equivalent to the expected value in non-moral decision making (i.e., how to optimize the number of lives) have yet to be elaborated. Third, we may struggle with which moral principles to follow and how to resolve the conflicts between principles—unsure preferences, which also refers to the uncertainty in our own subjective preference. Given the conflict between deontological and utilitarian demands, i.e., treating every single person equally or optimizing the number of lives that can be saved will always leave part of us unsatisfied. Forth, compared to classical binary moral dilemmas, this mineshaft problem provides us with three alternatives, with which how people distribute their mental effort into assessing each alternative is still an open question for moral psychologists (see Section 5.1.3). People may try to avoid ambiguity in Gate 1 and 2 in this situation, so Gate 3 is more likely to be selected. Or people may

[16] One case is slightly modified version of a case discussed in Parfit (2011).

adopt moral heuristic strategies, such as assigning equal probability to both Shaft A and B in which the 100 miners are trapped, and compute expected moral values of all three alternatives. In this case, Gate 3 has the highest expected moral value, so again Gate 3 is more likely to be selected. Besides, the number of miners that can be saved in Gate 3 can be manipulated to be greater, equal to, or less than the expected value of Gate 1 or 2. With the manipulation, how people balance outcomes (number of lives that can be saved) and their likelihoods (known or unknown) and how they estimate unknown events under different conditions are worth exploring in future studies. Thus, to examine whether we can generate the experimental findings based on two-forced-alternative-choice to multi-alternative/multi-attribute is essential to better understand the mechanism of moral decision-making in a real-world setting.

Moral decisions with multiple alternatives/attributes

Moral situations we encounter every day usually have more than two options; the examination of moral decisions with multiple alternatives/attributes is worthwhile revealing the real underpinnings of moral decisions. This section attempts to point out several necessary steps of bridging the gap between moral decisions and multialternative/multiattribute non-moral decision theories. First, we need to reconstruct moral situations with multiple options, instead of only two options. For each option, probabilities of its outcomes may be utilized to reconcile with people's counterfactual thoughts (e.g., it is impossible for the large man to stop the trolley). Also, these moral situations would not be identifiable, as the replications of similar moral situations should be experimentally independent. Second, the assessment of each attribute of all options needs to be quantified, which can be pre-determined or collected in the experiment by a piece-by-piece presentation. Note this quantification may not be restricted to utilitarianism, as different attributes may be estimated based on different moral principles, such as deontological rules and causal structure theory. Third, with more than two alternatives, context effects may be elicited in moral decisions. The similarity effect has been reported with moral trilemmas (Shallow et al., 2011). Whether the attraction and compromise effects would be seen in multialternative moral decisions has yet to be studied. These context effects may shed light upon the mechanisms of moral decisions in various moral contexts.

Consider a multi-alternative variant of the trolley problem (modified from Shallow et al., 2011): a trolley is speeding out of control threatening five railway workers, in which three options are presented:

1. Push intervention (P): Pushing a large man off the bridge, who can stop the trolley with probability p_p to save the five workers. If the trolley remains in motion, then all the six people will die.

2. Switch intervention (S): Flipping a switch, which will be in effect with probability p_S. If the switch works, then the trolley will be redirected to a side track to kill two innocent people and the five workers will be saved; otherwise, the five will be killed.

3. Omission (O): Doing nothing, then the trolley will remain in motion with probability p_O. If so, then the five workers will die.

Each of the three interventions consists of three attributes: N- the number of lives lost, R- the success rate of interventions, and I- the way of intervention. Manipulating the similarity of three interventions through changing their attribute values may elicit different choice behavior.

Regarding quantitative measurement of attributes for each alternative, a selective set of codifiable and uncodifiable principles may be needed. Different moral principles—deontology, utilitarianism principles, and causal structure theory, as well as personality and culture may influence people's moral subjective assessment of each attribute to varying degrees. For example, people would favor harmful omission according to omission bias but would disapprove of any harmful behavior according to deontological rules (see Chapter 1). The subjective evaluation of attribute N may be related to utilitarianism. The subjective evaluation of attribute R may involve people's risk attitudes. The measurement of attribute I may reflect the intensity of action aversion. It would be interesting to compare how people perceive the outcomes and their likelihood under moral situations with those under non-moral situations, like gambling. In non-moral situations people often exhibit loss aversion and the perception or transformation of objective probabilities may not be linear. For example, in CPT the probability weighting function indicates people tend to overestimate low probabilities but underestimate high probabilities, and (see Section 4.1)

To capture the relative evaluation of an attribute of all alternative, the notion of "loss aversion[17]" in non-moral decisions may need to be generalized in moral decisions. Loss aversion can be related to action and/or affective intuition. Action related loss aversion may be evoked toward alternatives P and S compared with O, according to deontological rules (do not harm others) and omission bias (indirect action preference). Similarly, direct physical contact involved in P may also elicit physical proximity related aversion compared to S. Affective intuition related aversion may be evoked by losing more lives (negative outcomes) or by conducting harmful action (negative action). Loss aversion may thus be resulted from both codifiable and uncodifiable moral principles. The next step would be to integrate evaluation of each attribute in the given moral context based on both codifiable and uncodifiable moral principles to predict moral decisions.

During the deliberate time, people may shift back and forth favoring one option and another at next moment. Some interesting questions we may ask are whether there are individual differences in the pattern of attention fluctuation (e.g., attention bias at attributes), and whether we can manipulate attributes (e.g., salience) of alternatives so that people's moral decisions can be altered. The LCA, the MDFT, and the MLBA might account for attention fluctuation during the processes of multialternative/multiattribute moral decision-making. The attention fluctuation between attributes can be influenced by evaluation from several codifiable and/or uncodifiable principles, such as deontological rules, utilitarianism, causal structure theory, and affective arousal, and/or moral heuristics. For instance, people may pay more attention on the intervention attribute if they are more committed to deontology. In particular, different emotional reaction may enhance or diminish the weights of attention shifts assigned to each attribute under consideration.

Meanwhile, model parameters estimated from actual data of each individual may help to explore individual differences. But note the implementation of models requires a reasonable large dataset collected from one subject, in order to estimate parameters of the models. In order to obtain the measurements of similar moral situations many times from a single subject, we may

[17] Trueblood et al. (2014) suggested using "disadvantage aversion" instead, as the three context effects have been observed in inference and perceptual experiment. In these experiments, the attributes are not hedonic, so that there is no losses or gains. But it is still premature to replace loss aversion with "disadvantage aversion".

need to convert moral dilemmas in current text format into presentations which can be repeated. Moral situations can be presented through visualization or animation with simple moral contexts (see Section 5.2).

Moreover, we may manipulate experiment conditions, such as imposing time pressure, to further examine the mechanisms of moral decisions. With estimated parameters of models, like the MLBA, we can compare whether certain parameter would differ in different conditions. Under time pressure, for instance, people may have different patterns of attention fluctuation (e.g., more bias toward one attribute) compared with no time pressure situations.

Moral situations can be much complicated in the real world that we have to take into account not only multiple alternatives but also multiple moral principles and the interaction between moral agents and the environment. Moral principles may disagree with each other and the environment may play a crucial role of shaping the processes of moral decisions. Thus, we need to address which moral principles would be considered, how to resolve their potential conflicts, and how to integrate the impact of the environment into specific moral decisions. Moral heuristics may provide a feasible approach when faced with moral situations with multiple alternatives and/or attributes. For example, in the above multi-alternative trolley problem, some moral principles may be ignored or all relevant attributes are assigned with equal weights for further evaluation. Therefore, to specify the role of moral heuristics in multi-alternative/multi-attribute moral decisions is an indispensable step to understand the processing of real-world moral decisions.

Dynamics of moral decisions

The last question is how we make moral decisions in the dynamic environment. In the real world, information is changing and may be updated at times, how do we incorporate the most recent information into the current state of mental computations in order to make optimal choices or reasonable moral judgments?

Virtue reality has been used to study moral actions in the classical moral dilemmas (Iliev et al., 2012; Navarrete et al., 2012; Patil et al., 2013; Skulmowski et al., 2014), whose results can inspire future experiments. First, emotional arousal was greater only in the case of “action” as compared with “omission” (Navarrete et al., 2012). Second, the discrepancy of virtual actions

and abstract judgments inherited order effects (Patil et al., 2013). Third, the virtual setting enables researchers to measure physiological and biological responses at the same time during moral actions being taken, such as electrodermal activity (Patil et al., 2013), pupillometric measurements (Skulmowski et al., 2014), eye tracking (Brandstätter & Körner, 2014). Specifically, eye tracking technology would help to reveal the actual information search strategies (i.e., how people examine available information) during moral decision-making. The additional evidence pointed a relatively new direction in research of moral decision, which hopefully would deepen understanding of dynamics of moral judgment.

Nevertheless, in the virtual reality settings, multiple repetition of similar moral problems cannot be tested in a single subject. The limited data from one subject might be insufficient for finer examination, such as cognitive computational approaches. A solution is to convert current moral problems into simple animation, with which seemingly similar moral problems but under various conditions many trials can be repeated within one subject. The feasibility of the animation presentation relies on that people may associate geometric shapes with social characters (Heider & Simmerl, 1944) and then act based on the virtual social interaction. The animation presentation of moral problems may disclose moral decisions from a novel perspective.

Further, with the emergence of autonomous vehicles, car manufacturers are facing a pressing issue about how to implement moral principles with concrete algorithms when the autonomous cars have to decide between running pedestrians and sacrificing themselves and their passengers to save the pedestrians. This social driverless dilemma (Bonnefon, Shariff, & Rahwan, 2016; Greene, 2016) places a formidable challenge in front of philosophers, policy regulators, and politicians, as it has yet to explore how to design autonomous machines that comport with our moral sensitivities. Therefore, to study moral decisions in the real world has been becoming more and more necessary for both theoretical and practical purposes.

REFERENCES

Allais, M. (1990). Allais Paradox. In *Utility and Probability* (pp. 3–9). London: Palgrave Macmillan UK. http://doi.org/10.1007/978-1-349-20568-4_2

Alós-Ferrer, C. (2016). A Dual-Process Diffusion Model. *Journal of Behavioral Decision Making*, 1–19.

Ayars, A. (2016). Can model-free reinforcement learning explain deontological moral judgments? *Cognition*, *150*, 232–242. http://doi.org/10.1016/j.cognition.2016.02.002

Baron, J. (2008). Moral judgment and choice. In *Thinking and Deciding* (4 ed.). Cambridge, UK.

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

Baron, J., & Ritov, I. (2009). Protected Values and Omission Bias as Deontological Judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making The psychology of learning and motivation* (Vol. 50, pp. 133–167). San Diego, CA: Elsevier Inc. http://doi.org/10.1016/S0079-7421(08)00404-0

Baron, J., & Spranca, M. (1997). Protected Values. *Organizational Behavior and Human Decision Processes*, *70*(1), 1–16.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*(2), 381–417. http://doi.org/10.1016/j.cognition.2008.03.001

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2014). Moral judgment and decision making. In K. Keren & G. Wu (Eds.), *Blackwell Reader of Judgment and Decision Making*. Malden, MA.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*(1), 154–161. http://doi.org/10.1016/j.cognition.2011.05.010

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting External Validity: Concerns about Trolley Problems and Other Sacrificial Dilemmas in Moral Psychology. *Social and Personality Psychology Compass*, *8*(9), 536–554. http://doi.org/10.1111/spc3.12131

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The Costs and Benefits of Calculation and Moral Rules. *Perspectives on Psychological Science*, *5*(2), 187–202. http://doi.org/10.1177/1745691610362354

Bersoff, D. M., & Miller, J. G. (1993). Culture, Context, and the Development of Moral Accountability Judgments. *Developmental Psychology*, *29*(4), 664–676.

Białek, M., & De Neys, W. (2017). Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity, *12*(2), 148–167.

Bonnefon, J.-F., Shariff, A. F., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1570–1573. http://doi.org/10.1126/science.aaf2729

Brandstätter, E., & Körner, C. (2014). Attention in risky choice. *Actpsy*, *152*(C), 166–176. http://doi.org/10.1016/j.actpsy.2014.08.008

Byrne, R. M. J. (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*(1), 135–157. http://doi.org/10.1146/annurev-psych-122414-033249

Coombs, C. H., Dawes, R. M., & Tversky, A. (1970). Mathematical Psychology: an Elementary Introduction. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366. http://doi.org/10.1016/j.tics.2013.06.005

Cushman, F. (2016). The Psychological Origins of the Doctrine of Double Effect. *Criminal Law and*

Philosophy, 1–14. http://doi.org/10.1007/s11572-014-9334-1
Cushman, F., & Young, L. (2011). Patterns of Moral Judgment Derive From Nonmoral Psychological Representations. *Cognitive Science*, *35*(6), 1052–1075. http://doi.org/10.1111/j.1551-6709.2010.01167.x
Cushman, F., Young, L., & Greene, J. D. (2009). *Our multi-system moral psychology: Towards a consensus view* (pp. 1–20).
Cushman, F., Young, L., & Hauser, M. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgment. *Psychological Science*, *17*(12), 1082–1089. http://doi.org/10.1111/j.1467-9280.2006.01834.x
Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*.
Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective & Behavioral Neuroscience*, *14*(2), 473–492. http://doi.org/10.3758/s13415-014-0277-8
Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. http://doi.org/10.1016/j.conb.2008.08.003
Dwyer, S. (2009). Moral Dumbfounding and the Linguistic Analogy: Methodological Implications for the Study of Moral Judgment. *Mind Language*, *24*(3), 274–296.
Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics*, *75*(4), 643–669. http://doi.org/10.2307/1884324
Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. http://doi.org/10.1146/annurev.psych.59.103006.093629
Farmer, G. D., Warren, P. A., El-Deredy, W., & Howes, A. (2016). The Effect of Expected Value on Attraction Effect Preference Reversals. *Journal of Behavioral Decision Making*, *63*(4), 223–9. http://doi.org/10.1002/bdm.2001
Fehr, E., & Fischbacher, U. (2004). Third Party Sanctions and Social Norms. *Evolution and Human Behavior*, *25*, 63–87.
Fleischhut, N. (2013). Moral Judgment and Decision Making under Uncertainty (pp. 1–135).
Foot, P. (2002). Moral Dilemmas. Oxford University Press on Demand.
Gigerenzer, G. (2010). Moral Satisficing: Rethinking Moral Behavior as Bounded Rationality. *Topics in Cognitive Science*, *2*(3), 528–554. http://doi.org/10.1111/j.1756-8765.2010.01094.x
Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. http://doi.org/10.1111/j.1756-8765.2008.01006.x
Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. http://doi.org/10.1146/annurev-psych-120709-145346
Gong, H., & Medin, D. (2012). Construal levels and moral judgment: Some complications. *Judgment and Decision Making*, *7*(5), 628–638.
Goodwin, G., & Darley, J. M. (2010) The perceived objectivity of ethical beliefs: psychological findings and implications for public policy. *Review of Philosophy and Psychology*, 1, 161-188.
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. http://doi.org/10.1037/a0015141
Greene, J. D. (2007a). The secret joke of Kant's soul. In *Moral psychology* (pp. 35–80). Cambridge, MA: MIT Press.
Greene, J. D. (2007b). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323.

http://doi.org/10.1016/j.tics.2007.06.004
Greene, J. D. (2016). *Our driverless dilemma*. *Science*, *352*(6293), 1–3.
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. http://doi.org/10.1016/j.cognition.2007.11.004
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, *44*(2), 389–400. http://doi.org/10.1016/j.neuron.2004.09.027
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, *293*(5537), 2105–2108. http://doi.org/10.1126/science.1062872
Guo, L., Trueblood, J. S., & Diederich, A. (2015). A Dual-process Model of Framing Effects in Risky Choice (pp. 1–6). Presented at the 37th Annual Conference of the Cognitive Science Society, Austin, TX.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. http://doi.org/10.1037/0033-295X.108.4.814
Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science*, *316*(5827), 998–1002. http://doi.org/10.1126/science.1137651
Haidt, J., & Joseph, C. (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturally Variable Virtues. *Daedalus*, *133*, 55–66.
Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral Dumbfounding: When Intuition Finds No Reason. *Lund Psychological Reports*, *2*, 1–23.
Hauser, M., Cushman, F., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A Dissociation Between Moral Judgments and Justifications. *Mind Language*, *22*(1), 1–21.
Heider, F., & Simmerl, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259.
Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 1–7. http://doi.org/10.1016/j.tics.2009.09.004
Hertwig, R., & Herzog, S. M. (2009). Fast and frugal heuristics: tools of social rationality. *Social Cognition*, *27*(5), 661–698.
Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2008). Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*, *15*(8), 534–539.
Holyoak, K. J., & Powell, D. (2016). Deontological coherence: A framework for commonsense moral reasoning. *Psychological Bulletin*, *142*(11), 1179–1203. http://doi.org/10.1037/bul0000075
Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science*, *310*(5754), 1678–1680. http://doi.org/10.1126/science.1120640
Huber, J., Payne, J. W., & Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research*, *9*(1), 90–98. http://doi.org/10.1086/208899
Huebner, B., Dwyer, S., & Hauser, M. (2008). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, *13*(1), 1–6. http://doi.org/10.1016/j.tics.2008.09.006
Huettel, S. A., Stowe, C. J., Gordon, E. M., Warner, B. T., & Platt, M. L. (2006). Neural Signatures of Economic Preferences for Risk and Ambiguity. *Neuron*, *49*(5), 765–775. http://doi.org/10.1016/j.neuron.2006.01.024
Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition*, *40*(8), 1387–1401. http://doi.org/10.3758/s13421-012-0217-1

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, *93*(5), 1449–1475.
Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 267–294). Cambridge, UK.
Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263–292.
Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the Assumptions of Economics. *The Journal of Business*, *59*(S4), S285–S330. http://doi.org/10.1086/296367
Kant, I. (1996). The Metaphysics of Morals. (M. Gregor, Ed.). Cambridge University Press.
Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(279), 190–194. http://doi.org/10.1111/1467-8284.00419
Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis*, *64*(282), 181–187. http://doi.org/10.1111/j.1467-8284.2004.00481.x
Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications*, *6*, 7455–9. http://doi.org/10.1038/ncomms8455
Loewenstein, G., O'Donoghue, T., & Bhatia, S. (2015). Modeling the interplay between affect and deliberation. *Decision*, *2*(2), 55–81. http://doi.org/10.1037/dec0000029
Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*(C), 133–143. http://doi.org/10.1016/j.cognition.2015.11.008
McDowell, J. H. (1998). Mind, Value, and Reality. Harvard University Press.
Metcalfe, J., & Mischel, W. (1991). A hot/cool-system analysis od delay of grafication: dynamics of willpower. *Psychological Review*, *106*(1), 3–19.
Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. http://doi.org/10.1016/j.tics.2006.12.007
Miller, J. G., & Bersoff, D. M. (1992). Culture and Moral Judgment: How Are Conflicts Between Justice and Interpersonal Responsibilities Resolved? *Attitude and Social Cognition*, *62*(4), 541–554.
Miller, R., & Cushman, F. (2013). Aversive for Me, Wrong for You: First-person Behavioral Aversions Underlie the Moral Condemnation of Harm. *Social and Personality Psychology Compass*, *7*(10), 707–718. http://doi.org/10.1111/spc3.12066
Moran, L. (2015, June 1). Ohio woman must walk 30 miles — same distance of taxi fare she refused to pay: judge. Retrieved May 9, 2017, from http://www.nydailynews.com/news/national/judge-orders-fab-fare-skipping-ohio-woman-walk-30-miles-article-1.2242266
Mukherjee, K. (2010). A dual system model of preferences under risk. *Psychological Review*, *117*(1), 243–255. http://doi.org/10.1037/a0017884
Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem." *Emotion*, *12*(2), 364–370. http://doi.org/10.1037/a0025561
Neumann, von, J., & Morgenstern, O. (1947). Theory of Games and Economic Behavior (60 ed.). Princeton University Press.
Nichols, S. (2002). Norms with feeling: Towards a psychological account of moral judgment. *Cognition*, *84*, 221–236.
Nichols, S. (2004). Sentimental Rules.
Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530–542. http://doi.org/10.1016/j.cognition.2005.07.005
O'Hara, R. E., Sinnott-Armstrong, W., & Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments, 1–8.

Parfit, D. (2011). On What Matters. Oxford, England: Oxford University Press.
Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., Silani, G., Patil, I., et al. (2013). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social Neuroscience*, *9*(1), 94–107. http://doi.org/10.1080/17470919.2013.870091
Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, *112*(13), 4170–4175. http://doi.org/10.1073/pnas.1415250112
Petrinovich, L., & ONeill, P. (1996). Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobioiogy*, *17*, 145–171.
Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: A comment on Haidt (2001). *Psychological Review*, *110*(1), 193–196. http://doi.org/10.1037/0033-295X.110.1.193
Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, *9*(1), 29–43. http://doi.org/10.1080/13869790500492466
Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75. http://doi.org/10.1037/a0021867
Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.
Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370–392. http://doi.org/10.1037//0033-295X.108.2.370
Royzman, E. B., & Baron, J. (2002). The Preference for Indirect Harm. *Social Justice Research*, *15*(2), 165–184.
Royzman, E. B., Leeman, R. F., & Baron, J. (2009). Unsentimental ethics: Towards a content-specific account of the moral–conventional distinction. *Cognition*, *112*(1), 159–174. http://doi.org/10.1016/j.cognition.2009.04.004
Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Publishing Group*, *15*(8), 1–14. http://doi.org/10.1038/nrn3776
Savage, L. (1954). The foundations of statistics. NY.
Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as Embodied Moral Judgment, *34*(8), 1096–1109. http://doi.org/10.1177/0146167208317771
Schwitzgebel, E., & Cushman, F. (2012). Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind Language*, *27*(2), 135–153.
Scott, H. P. (2006). Cumulative Prospect Theory's Functional Menagerie. *Journal of Risk and Uncertainty*, *32*, 101–130.
Shallow, C., Iliev, R., & Medin, D. (2011). Trolley problems in context, *6*(7), 593–601.
Shenhav, A., & Greene, J. D. (2010). Moral Judgments Recruit Domain-General Valuation Mechanisms to Integrate Representations of Probability and Magnitude. *Neuron*, *67*(4), 667–677. http://doi.org/10.1016/j.neuron.2010.07.020
Simon, D., Stenstrom, D. M., & Read, S. J. (2015). The coherence effect: Blending cold and hot cognitions. *Journal of Personality and Social Psychology*, *109*(3), 369–394. http://doi.org/10.1037/pspa0000029
Simon, H. A. (1957). Models of man. New York: Wiley.
Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*, 1–19.
Simonson, I. (1989). Choice Based on Reasons: The Case of Attraction and Compromise Effects. *Journal of Consumer Research*, *16*(2), 158–174. http://doi.org/10.1086/209205
Skulmowski, A., Bunge, A., Kaspar, K., & Pipa, G. (2014). Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. *Frontiers in Behavioral Neuroscience*, *8*, 60. http://doi.org/10.3389/fnbeh.2014.00426
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*,

119(1), 3–22.
Smart, J. J. C., & Williams, B. (1973). Utilitarianism. Cambridge: Cambridge University Press. http://doi.org/10.1017/CBO9780511840852
Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Jounral of Experimental Social Psychology*, *27*(1), 76–105.
Sripada, C. S., & Stich, S. (2006). A Framework for the Psychology of Norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *Innateness and the Structure of the Mind* (Vol. II, pp. 280–301). New York.
Sunstein, C. R. (2003). Moral Heuristics and Moral Framing Lecture. *Minnesota Law Review*, *88*, 1556–1597.
Sunstein, C. R. (2005). Moral Heuristics. *Behavioral and Brain Sciences*, *28*, 531–573. http://doi.org/10.1017/S0140525X05000099
Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*(3), 454–458. http://doi.org/10.1016/j.cognition.2011.01.018
Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.
Thurstone, L. L. (1927). A law of comparative judgment, 273–286.
Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*(2), 440–463. http://doi.org/10.1037/a0018963
Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). Multiattribute Linear Ballistic Accumulator Model of Context Effects in Multialternative Choice. *Psychological Review*, *121*(2), 179–205. http://doi.org/10.1037/a0036137.supp
Trueblood, J. S., Brown, S. D., Heathcote, A., & Busemeyer, J. R. (2013). Not Just for Consumers. *Psychological Science*, *24*(6), 901–908. http://doi.org/10.1177/0956797612464241
Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*(4), 281–299.
Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.
Tversky, A., & Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*, *211*(30), 453–457.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. http://doi.org/10.1007/BF00122574
Tversky, A., Slovic, P., & Kahneman, D. (1990). The Causes of Preference Reversal. *The American Economic Review*, *80*(1), 204–217. http://doi.org/10.2307/2006743
Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.
Usher, M., & McClelland, J. L. (2004). Loss Aversion and Inhibition in Dynamical Models of Multialternative Choice. *Psychological Review*, *111*(3), 757–769. http://doi.org/10.1037/0033-295X.111.3.757
Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100. http://doi.org/10.1016/j.cognition.2010.04.003
Valdesolo, P., & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science*, *17*(6), 476–477. http://doi.org/10.1111/j.1467-9280.2006.01731.x
Van Bavel, J. J., Jenny Xiao, Y., & Cunningham, W. A. (2012). Evaluation is a Dynamic Process: Moving Beyond Dual System Models. *Social and Personality Psychology Compass*, *6*(6), 438–454. http://doi.org/10.1111/j.1751-9004.2012.00438.x
Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a Bomb on a Person Versus Throwing a Person on a Bomb. *Psychological Science*, *18*(3), 247–253. http://doi.org/10.1111/j.1467-9280.2007.01884.x

Waldmann, M. R., & Wiegmann, A. (2010). A Double Causal Contrast Theory of Moral Intuitions in Trolley Dilemmas (pp. 2589–2594). Presented at the 32nd annual conference of the cognitive science society.

Wedell, D. H. (1991). Distinguishing Among Models of Contextually Induced Preference Reversals. *Journal of Experimental Psychology*, *17*(4), 767–778.

Wheatley, T., & Haidt, J. (2005). Hypnotic Disgust Makes Moral Judgments More Severe. *Psychological Science*, *16*(10). http://doi.org/10.1111/j.1467-9280.2005.01614.x

Wiegmann, A., Okan, Y., & Nagel, J. (2011). Order Effects in Moral Judgments. *Philosophical Psychology*, 1–38.

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, *47*(3), 276–332. http://doi.org/10.1016/S0010-0285(03)00036-7

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the Neural and Cognitive Basis of Moral Luck: It's Not What You Do but What You Know. *Review of Philosophy and Psychology*, *1*(3), 333–349. http://doi.org/10.1007/s13164-010-0027-y