



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

“Virtual Reality for Investigating Moral Decision-Making”

verfasst von / submitted by

Piotr Mateusz Patrzyk, Lic.

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

A 066 013

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Joint Degree Programme MEi:CogSci Cognitive Science

Betreut von / Supervisor:

Univ.-Lektor Dipl.-Ing. Dr. Paolo Petta, OFAI

Acknowledgements

I would like to thank my supervisor, Paolo Petta, for helping me define the goal of this work and believing in its value, in spite of encountered difficulties.

Contents

Abstract	v
Zusammenfassung	vii
List of Tables	viii
List of Figures	ix
1 Introduction	1
2 Moral Faculty	3
2.1 Aggressive Morality	4
2.2 Defensive Morality	7
2.3 Integration	10
3 Investigating Morality	15
3.1 Vignettes	16
3.2 Behavioral Game Theory	20
3.3 Virtual Reality	24
4 The Counterfeiting Game	27
4.1 Problem Introduction	27
4.2 Description of the Game	28
4.3 Evaluation of the Game	33
5 Discussion	35
Bibliography	40
A Game Scenario	57
A.1 Text-based Description	57
A.2 Game Script	59
Curriculum Vitae	61

Abstract

Investigating human moral faculty has proven a difficult research endeavor. Due to differences between what people actually do and what they want others to think they do, whenever moral considerations are involved, self-reports of moral behavior are unreliable. Experimental expectancy effects and self-deceptive illusions held by people effectively prevent them from reporting accurate facts and motivations when inquired. Therefore, indirect methods of studying human moral intuitions have been developed. The most widely used method consists of presenting subjects with written vignettes and asking them what they or other story characters would or should do. Unfortunately, the dilemmas described in these texts are very often unusual or altogether unrealistic, what leads to frivolous treatment by subjects. Distinctive features of situations involving moral considerations are the environment itself being risky or characters operating under pressure. Virtual Reality (VR) has been argued to be a panacea to recreate and control such real-world situations for research purposes, by referring to relative ease of manipulating the environment.

To critically assess this approach, we prototype a non-immersive desktop-based VR system using the Unity Game Engine. Within this 3D environment, the user controls a character from a first-person perspective. The scenario is set in a company producing car accessories, and the player needs to decide whether to engage in production of counterfeit goods. Forced interactions with non-player characters are aimed at revealing the player's appraisals and mental representations of the situation, as well as the justification of their moral decision.

This pilot study thus sub-serves the development of a grounded critical understanding of theoretical, empirical, and methodological issues in the design and implementation of such a scientific VR-based probe. We critically assess our prototype and explicate key considerations for the interdisciplinary development and deployment of VR-based research methods in decision-making in social contexts. In particular, we address the different available support for modeling and simulating of physical and social aspects, as well as the required efforts and competences. In this way, we aim to contribute to a realistic, and thereby productive and responsible attitude in the field.

Zusammenfassung

Die Untersuchung der menschlichen moralischen Fähigkeiten hat sich als schwieriges Forschungsvorhaben herausgestellt. Auf Grund der Unterschiede zwischen tatsächlichem Tun und der gewünschten Vorstellung anderer über das eigene Tun sind, sobald moralische Erwägungen involviert sind, Selbst-Berichte über moralisches Verhalten unzuverlässig. Experimentelle Erwartungseffekte und selbstbetrügerisches Wunschenken verhindern eine korrekte Darstellung von Fakten und Motivationen auf Befragung. Aus diesem Grund wurden indirekte Methoden für das Studium menschlicher moralischer Intuitionen entwickelt. Die weitverbreitetste Methode besteht aus darin, ProbandInnen schriftliche Vignette vorzulegen und diese zu befragen, was sie oder andere Figuren der Geschichte tun würden oder sollten. Unglücklicherweise sind die in diesen Texten beschriebenen Dilemmas sehr oft unüblich oder vollkommen unrealistisch, was zu leichtfertiger Auseinandersetzung durch die ProbandInnen führt. Situationen, die moralische Erwägungen einschließen, sind durch risikoreiche Umgebungen oder unter Druck handelnde Figuren gekennzeichnet. Virtual Reality (VR) ist als Ideallösung für die Nachbildung und kontrollierte Steuerung solcher realer Situationen für Forschungszwecke befürwortet worden, unter Verweis auf die verhältnismäßig einfache Manipulierbarkeit der Umgebung.

Für eine kritische Beurteilung dieses Ansatzes entwickelten wir einen Prototyp eines nicht-immersiven Desktop VR Systems unter Verwendung der Unity Game Engine. In dieser 3D Umgebung steuert die BenutzerIn eine Figur aus der egozentrischen Perspektive. Das Szenario handelt von einer Firma, die Autozubehör produziert, und die SpielerIn muss entscheiden, ob in die Produktion von gefälschten Waren eingestiegen werden soll. Obligatorische Interaktionen mit simulations-gesteuerten Figuren zielen auf die Aufdeckung der subjektive Einschätzungen und mentalen Repräsentationen der Situation durch die SpielerIn, sowie der Rechtfertigung ihrer moralischen Entscheidung.

Diese Pilotstudie dient der Entwicklung eines verankerten kritischen Verständnisses theoretischer, empirischer, und methodologischer Problempunkte in Entwurf und Implementierung eines solchen VR-gestützten Untersuchungsmittels. Gestützt auf der kritischen Beurteilung unseres Prototyps erläutern wir wesentliche Überlegungen zur interdisziplinären Entwicklung

und dem Einsatz von VR-gestützten Forschungsmethoden zur Entscheidungsfindung in sozialen Kontexten. Dabei streichen wir insbesondere die unterschiedlichen Umfänge der verfügbaren Unterstützung zur Modellierung und Simulation physischer und sozialer Aspekte, sowie die erforderlichen Aufwände und Kompetenzen heraus. Damit soll diese Arbeit einen Beitrag für eine realistische und damit produktive und verantwortungsvoller Haltung liefern.

List of Tables

- 2.1 Justification Strategies 9
- 2.2 Functions of Morality 10
- 2.3 Components of the Moral Faculty 13

- 3.1 Prisoner’s Dilemma 21
- 3.2 Summary of Research Methods 26

- 4.1 Justifications for Moral Actions 33

List of Figures

- 2.1 Strategic Interaction in Morality 8
- 2.2 The Uses of Moral Faculty 13

- 3.1 Example Variation of the Trolley Dilemma 18

- 4.1 Opening Scene of the Counterfeiting Game 31
- 4.2 The Company Building 31
- 4.3 Within-Game Conversation 32

- 5.1 Effort-Validity Tradeoff 37

- A.1 First Conversation of the Game 59
- A.2 Second Conversation of the Game 60

Chapter 1

Introduction

In this thesis we describe and critically evaluate the use of *Virtual Reality* (VR) as a research tool in moral psychology. Moral judgment and behavior are social phenomena in dynamic interactions. They are characterized by social influence on responses individuals give. For this reason, researchers need to develop tools that reveal the nature of these influences and how they impact obtained results.

Virtual Reality has been proposed to serve as a research tool in social psychology because it provides *psychologically* realistic, dynamic, and multi-modal experience of an environment. In such a way, it overcomes limitations of text-based and game-theoretic methods that are unable to realistically represent social contexts in which decisions are made.

As research in moral psychology has extensively relied on decontextualized text-based stimuli, researchers have become aware of their limitations and called for development of valid methods. Coming from this tradition, we have undertaken the effort in developing a prototype of a new method. The choice of Virtual Reality has been motivated by the availability of tools that can be employed to develop VR-based research instruments and the promise of viability of such endeavor, due to abundance of learning materials in the Internet and the existence of a strong user community.

We report our experience with developing a concrete research question to be addressed in a Virtual Reality simulation and prototyping an actual implementation with commodity off-the-shelf tools. Based on our hands-on experience that enabled us to contextualize, validate, and complement the related literature, we point out several caveats in the interdisciplinary method development in the context of investigating social decision-making.

In Chapter 2 we introduce human moral faculty. We identify and discuss the social situations it is adapted to and what underlying mechanisms have been proposed. We identify three broad areas in which morality is employed: The first is *aggressive* morality, which we discuss

in Section 2.1 — this component is responsible for attacking individuals performing actions undesirable for the actor; we discuss how individuals generate and agree on types of actions that are *moralized* and how this process can be investigated. In Section 2.2 we address *defensive* morality, which is employed in situations where an actor fears becoming a victim of moralistic aggression from others; we outline how individuals make decisions that can negatively impact interests of others and how such decisions can be investigated. In Section 2.3 we discuss how individuals make judgments that are supposed to be *impartial*; as no human judgment can *actually* be impartial, we cover influences on these decisions and how researchers should make them explicit.

In Chapter 3 we review research instruments to investigate human moral faculty. We review major tools and obtained findings. We present fundamental arguments against the use of self-reports to investigate human morality and then we review other strategies. In Section 3.1 we review the use of *written vignettes* in moral psychology and find that while they are very good at eliminating experimental expectancy effects, they are crucially lacking in situatedness, thereby failing to provide realistic stimuli. In Section 3.2 we cover *behavioral game theory* as a framework for modeling moral decisions. We explain how these instruments excel at clearly defining interests and roles of people participating in a moral decision but fail in experimental settings where people often do not interact with the instrument the way it is expected or give socially desirable responses. In Section 3.3 we introduce *Virtual Reality* as method of research in moral psychology and the main focus of this thesis. We explicate its potential to address the problems inherent in the previous methods.

In Chapter 4 we cover the research prototype developed by us. In Section 4.1 we introduce the problem of product counterfeiting and argue how it serves as a good example of moral decision that can be modeled in Virtual Reality. In Section 4.2 we detail the prototype instrument developed, and in Section 4.3 we report our experiences in building it and evaluate its strengths and weaknesses.

In Chapter 5 we present our methodological reflection on the process of developing Virtual Reality research tools for moral psychology and offer our insights for researchers considering the use of this method. We argue that while this method has its promises for serving as a research tool, its viability and actual contributions for the investigation of complex research questions have to be thoroughly assessed.

Chapter 2

Moral Faculty

What? A great man? I can only see an actor of his own ideal.

— Friedrich Nietzsche [1, p. 62]

Moral reasoning and decision-making are essential components of social cognition. We often perform actions that can have impact on other people's well-being. In the same way, we are affected by actions of others. In a world that is only partially observable, we can never be sure how others will behave toward us and what reactions our behavior will produce. Moreover, there might be non-foreseeable variability in how others interpret our actions. The fact that humans are interdependent creates a situation in which social interactions are strategic.

Moral faculty will be understood here as an evolved capacity that helps us navigate in the social world [2]. It is concerned with *judgments* and *behaviors* that respond to *conflicts of interests* among humans [3]. It helps us create *expectations* about how others will behave and what *consequences* our behavior may produce.

Throughout the thesis, the term *moral action* or *moral behavior* will denote any behavior committed in such a context, whereas *moral judgment* will denote an assertion made by human about a situation involving conflict of interests.

Four things need clarification in regard to such an approach:

1. This thesis is concerned with a *descriptive* approach to ethics [4]. For instance, if a husband found out that his wife cheated on him and killed her because of that, it would be a *moral behavior*. If he forgave her, it would also be a *moral behavior*. If a third person came in and said the act of killing was wrong, they would make a *moral judgment*. If another person came and said the act of killing was justified, they also would make a *moral judgment*. In the current thesis, I will have nothing to say about *normative* status of the discussed actions and judgments. Instead of that, my aim is to shed light on cognitive mechanisms

governing these.

2. The scope of conflicts of interests will be understood broadly. People differ not only in opinions concerning particular issues, but also at a deeper level regulating what they consider a moral domain in the first place [5]–[7]. Because of this diversity, moral decisions are often not concerned with the question *whether* to follow a rule or not, but rather *whose* rules to obey and *whose* rules to break. For instance, people constantly violate *someone’s* moral rules in their everyday decisions what to eat and how to dress. For this reason, I do not introduce any *a priori* distinction regarding the scope of morality and commit to a more functional approach that assumes everything can produce a conflict of interests, provided it is moralized [8].
3. *Moral faculty* will not be understood as a single, functionally isolated module. Such claim has been shown to be problematic [9], and it seems more appropriate to conceptualize morality as a capacity based on several non-moral adaptations [10], especially mental state attribution [11], [12] and dominance-seeking [13].
4. These adaptations will be discussed through the lens of evolutionary theory, which assumes that all moral phenomena — behavior, judgment, as well as mental representations and inference rules — are shaped through natural selection [14], [15] and thus are *ecologically rational*, meaning that they are overall adaptive solutions to problems posed by the organism’s environment [16]. Following terminology adopted in the field, by *ultimate* cause we will mean the ascribed evolutionary reason why a given behavior provides a competitive advantage, and by *proximate* cause we will mean the mechanism directly governing an individual’s behavior [17]. By the term *consciously realized* cause we will denote the reasons people think or report to be causes of their behavior, a concept fundamentally different from the first two [18]–[20].

In the remaining part of this chapter, we discuss the *functions* of moral faculty. In Section 2.1 we introduce the *aggressive* component of moral faculty and argue that this is a primary way in which morality is used. In Section 2.2 we introduce a *defensive* component of moral faculty. In Section 2.3 we discuss the ways in which these components are *integrated* into moral judgments.

2.1 Aggressive Morality

It is a natural predisposition of humans to seek pleasure and avoid pain. If something in the environment is undesirable, we try to distance ourselves from it. When other people threaten our

resources or reputation, we need to protect ourselves. Being considered weak by the group is not in a social organism's interest. Therefore, upon being challenged we feel anger and counterattack competitors [21], which sometimes leads to a violent escalation of a conflict (e.g. [22]).

The interesting thing about human morality is that people maintain complex rules over when it is accepted to attack others in response to things they did in the past. People would say that killing others only because they threatened one's reputation or resources need not be the right thing to do. The question *why* is far from trivial. In addition, the story can quickly become more complex, if we add further details: was the act of killing committed in a self-defense? Was it defense of one's family? Was the victim's attack intentional?

In tough environments where the competition for resources is essential for survival, the evolution favors aggressive traits causing organisms to fight even if they risk their lives (see [23]). But once an organism possesses higher cognitive functions allowing them to foresee the future, other considerations enter the picture. Killing other people is not always the best option, since the success of this action is not guaranteed. Since the other organism can fight back such decision is risky. We may always lose a fight. Moreover, since humans are attached to each other, even a successful act of killing might provoke later retaliation from other group members. When we deal with stronger organisms that can damage us, we need to plan strategically.

Additionally, irreversible elimination of another organism is not always the most beneficial outcome to the aggressor. This is certainly never the case with cockroaches, but very often with humans. As noted earlier, since humans are interdependent, we have a vested interest in how others will behave. If your partner does not behave in a way you want him to do, you can either end the relationship with them, thereby losing future benefits following from cooperation, or attempt to manipulate their behavior. Humans have evolved complex mental modules that allow them to evaluate the worth of given relationship and make decisions in such contexts [24], [25].

When humans are treated by others in an undesirable way, they attempt to change such behavior [26], [27]. This predisposition is primarily motivated by the proximate mechanism of anger [21], but also disgust and contempt [28], [29], which form a triad of moral emotions. One option in the case of conflict is physical coercion, which suffers from similar problems as the elimination attempts discussed before. Assuming equal motivation to obtain resources, this option is only open to the stronger individuals who can effectively threaten the weaker ones (see [30]). Another possibility is social coercion, which aims to enforce a given behavior without resorting to physical violence. If one decides to follow this path, morality becomes indispensable.

Morality is a tool that allows enforcing a given course of action without exposing the true

reasons of its desirability. Moral rules have been conceptualized as mere prescriptions beneficial to those who preach them [31]–[35]: The goal is to persuade the other person that following the course of action preferred by oneself is the decision they *should* take. If this course of action should coincide with their best interest, simple causal reasoning can be employed to prove that. But since interests of different people are very often in conflict and the action most beneficial to oneself need not be the most beneficial action to another. In such a case, one needs to perform a *moralistic manipulation* that attributes objective moral properties to actions [8], [36], [37]. If the other person then starts to recognize some intrinsic moral value of a given action, the manipulation can be considered successful.

Due to the social nature of humans, the process of invoking morality affects not only the direct target, but also the observers of an interaction. An attractive feature of moralistic aggression is that it not only avoids the risk of damage incurred in physical fight but may also be more successful in getting others to support the aggressor. Gossiping and slandering are defining characteristics of human social nature. When treated in an undesirable way, instead of a physical attack we often perform the more cost-effective strategy of threatening the target’s social reputation. In doing so, we attempt to invoke moral outrage in observers and recruit them to support us [38], [39]. This can be achieved both, after an undesirable behavior took place and us now wanting to exercise influence on a particular individual, and when we preventively preach moral rules to those who could potentially behave otherwise.

So far, we have considered the case in which humans moralize actions that can pose a direct threat to them. Another interesting feature of human morality is that very often we make judgments about people we have never interacted with. Claims that any rules we should preach be universal or that violators deserve the same punishment regardless of the victim’s identity, are typical for moralized actions [40]. Additionally, people who conspicuously broadcast moral rules and enforce behaviors deemed to be desirable in their groups gain respect from others [41], [42].

Why do we care about others who claim to be victims of (im)moral behaviors? While there are no direct reasons for joining condemnation campaigns against such perpetrators, there are indirect reasons why such behavior may be adaptive. It might happen that an observer’s welfare is related to the welfare of one side of the conflict, or that they are in some partnership relation [43]. It might also happen that all sides attempt to recruit third parties to fight for their case, which creates pressure to coordinate [44]–[46]. In such ways, moral cognition is closely related to coalition management [38]. Coordinating support based on action characteristics provides common decision rules that are efficient [45] and promote the fitness of rule-following

and coordinating agents [47].

Finally, it is important to note that the negative emotions and attitudes toward violators of our norms are not moral in a strict sense. For this hold, it would require us to feel outraged *because* we have seen a norm violation. In case when anger is provoked by some threat to our goals, our reaction is purely egoistic, since it is our personal welfare that is challenged. In the case of intervention as a third party, our reaction is altruistic, since it is then the welfare of another person we care about. The cases when we are motivated to fight against norm violations *per se* are not common in humans [48]. It needs to be emphasized that the predicate *moral* has a purely functional meaning in this context — people may often report anger due to norm violation but the ultimate and proximate reasons for it are in fact different (see also [49]).

Are there any constraints to making moral attacks on others? The emerging picture so far is that as long as we have social support, we can perform any unsupported moral accusations as we wish. In order to see rules can emerge from such interactions, it is necessary to first look at another component of moral cognition: our ability to defend ourselves.

2.2 Defensive Morality

Probably the most distinctive feature of moral decisions is the experience of internal struggle between what one *wants* to do and what one *should* do (see [50]). This is an interesting characteristic since it means we are considering foregoing opportunities to benefit ourselves (e.g., by stealing or raping) due to abstract concepts of norms, duties, and obligations.

An evolutionary reason why we conform to such kind of reasoning could be that we expect others to perform moralistic aggression toward us when we violate their interests. In a world where people are willing to attack us if we transgress their norms, it is risky to be exposed violating them. Thus, if we are considering committing an action that will likely be harmful to someone else we have the following options: (1) Refrain from committing this action, (2) conceal this action from others, or (3) justify this violation of interest.

Naturally, the first option entails a cost, as in that case we relinquish the benefit. The second option would be ideal, but we can never be sure whether our action actually remains undetected. For that reason, we need to engage in the strategic considerations inherent to the third option.

This need for behavior negotiation puts agents, patients, and observers in a *strategic* situation where they can influence each other (see Fig. 2.1). *Mutual influence* in this system is achieved through the existence of a justification component: it ensures that agents are able to defend themselves when challenged on the one hand, and that third parties need to have support for their accusation in the form of proof for moral norm violation on the other.

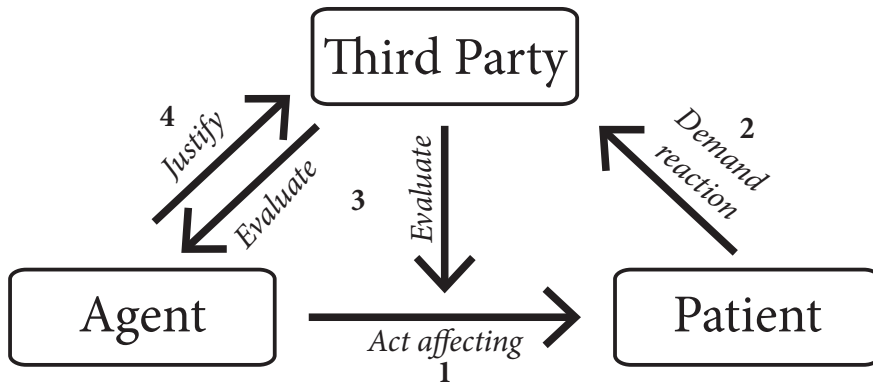


Figure 2.1: Strategic Interaction in Morality: When Agent takes an action affecting interests of Patient (1), Patient demands reaction (2), which is determined by evaluation (3) of the act and Agent, who is thus forced to justify their decision (4), introducing some constraint (see also [44], [51], for related analyses). This figure originally appeared in [52].

Indeed, guiding one’s own conduct and guiding the conduct of others are different tasks with different underlying motivations, so it is natural to treat them as distinct adaptations [53]. While in the case of aggressive use of morality rules were our friends, in the case of defending oneself rules can become a hindrance [48].

For this reason, defensive morality is primarily concerned with *dismissing* rules as not applicable to one’s own conduct. Humans are skilled in devising arguments in support of their desired conclusions [54]–[57], and moral reasoning is where this is employed probably most frequently.

The idea that before committing a given act humans cognitively reframe it for it to fit with expectations imposed by third parties was first put forward within criminology by Sykes and Matza [58] under the name *Neutralization Theory*. This process was later reinvented in psychology by Bandura [59]–[62] under the name *Moral Disengagement Theory* and has become more recognized in the recent research in this area [63]–[65]. Both approaches describe cognitive strategies that humans use in order to justify acts they are willing to commit. These strategies are summarized in Table 2.1.

Generating these justification serves two intertwined functions: (1) it helps us foreseeing others’ reaction and thus make an informed decision whether to take a risk, (2) it prepares us for defense against moralistic aggression.

The most exquisite way in which a harmful action can be conducted is through the use of moral justification (appeal to higher loyalties). In fact, the majority of violent and sadistic acts are committed for *higher purposes*, starting from ritual killing of animals and ending at large-

		<i>Moral Disengagement</i>	<i>Neutralization</i>
<i>Reframing</i>	conduct	Moral Justification	Appeal to Higher Loyalties
		Euphemistic Labeling	
		Advantageous Comparison	Denial of Injury
	effects	Distortion of Consequences	
	agency	Displacement of Responsibility	Denial of Responsibility
		Diffusion of Responsibility	
victim	Dehumanization	Denial of the Victim	
	Blaming the Victim		
authority		Condemnation of the Condemners	

Table 2.1: Comparison of the justification strategies described within *Moral Disengagement Theory* [59]–[62] and *Neutralization Theory* [58], as well as classifies them in regard to the target of cognitive reframing.

scale terrorist attacks (see [51]). In this strategy, a norm cited to be violated by is supposed to be canceled by a superior obligation of the actor.

In more casual cases where such appeal to higher loyalties is not promising, an agent might try to redefine an interest violation in a way that makes it harmless (*euphemistic labeling*), compare it more serious violation in order to weaken outrage (*advantageous comparison*), distort consequences, deny their own agency (devise external factors and circumstances supposed to attenuate blame), or argue that the interests of a victim do not deserve to be respected (due to provocation or need for retaliation). Finally, an agent might try to counter-attack the accusers (*condemnation of the condemners*) by exposing their spiteful motivation behind moralistic aggression.

There are some controversies in the literature whether these justification are produced before or after an action [66], [67]. It is probably most prudent to assume that it can happen at either times: when we consider future action and when we think about justifications for an action we have already committed [64], [65].

Action status negotiation is a proximate mechanism that helps an individual choose the optimal action. When an individual’s fitness depends on his reputation in the eyes of other

	<i>Aggressive</i>	<i>Defensive</i>
<i>Individual</i>	Enforce desired actions	Inhibit risky actions
<i>Social</i>	Motivate third parties	Justify interest violations

Table 2.2: Functions of Morality

group members, it is adaptive to maintain a good image. That gives birth to a sense of defensive morality in which individuals constrain their behavior due to the risk of long-term losses caused by broken social relationships. As Richard Alexander put it succinctly, conscience is *a still small voice that tells us how far we can go in serving our own interests without incurring intolerable risks or costs to our own interests* [3, p. 107].

Reputational concerns drive much (if not all) of ordinary moral behavior [68]–[71]. One important fact about human decision-making faculties needs to be noted at this point: since we take cognitive faculties of humans to be *ecologically rational* (or *bounded*), people may (and do) make behavioral mistakes. There are many cases where humans could behave in a more self-serving way, but follow the moral norm. Such behaviors can easily be misunderstood as manifestations of e.g. ultimately valuing moral rules for their own sake, or selfless prosocial concerns. In fact, this pattern of behavior can be seen to be caused by evolutionary pressure favoring cognitive biases when they increase an organism’s fitness overall [72]. In other words, because cognitive assets are constrained, we often make automatic decisions regarding norm violations. Since consequences of committing interest violations in situations where it would not be optimal (an agent could then be exposed and avenged) are worse than consequences of foregoing opportunities to benefit oneself, humans are biased toward automatic rule compliance [73], [74], producing less optimal behavior than a fully rational and omniscient agent.

In Section 2.3 I will discuss what patterns of human behavior follows from the integration of the functions of morality discussed so far.

2.3 Integration

So far I have reviewed the ways in which people use morality in interactions that involve them directly. In Section 2.1 I described the logic behind moralistic aggression and in Section 2.2 I described the ways in which people use moral principles to defend themselves. Findings of our literature review are summarized in Table 2.2.

A key problem with the complexity of moral cognition lies in the fact that these two functions

are integrated within one individual. For instance, performing moralistic aggression is useful, but it also has a drawback in the fact that preaching moral rules creates increased expectation in others to see the preacher following them [39]. Violating interests of others is also useful, but it creates a risk of retaliation if we do not have a good justification. Finally, people possess moral beliefs that are supposed to be *impartial*: where do they come from? Are they an outgrowth of *aggressive* or *defensive* morality?

Due to competing motives within an individual moral cognition is hypocritical by its nature — humans attempt to commit acts violating interests of others while trying to show them how moral they are [3], [48], [75]. They have biased perception of their own past actions and tell others stories with little or no connection to reality [19], [76]. At the same time, they will carefully inspect actions of other in order to evaluate their worth (see [77]).

Studies on cheating illustrate the problem with moral faculty. When people are given opportunity to cheat in laboratory studies in order to receive more money, they indeed do that, but most of them do not lie to the maximum extent possible, thereby avoiding unnecessary suspicions [78], [79]. Consider yourself in a following situation: you are given a dice and instructed to roll it privately. After you have rolled it come to me, tell me the result and I will pay you the amount of money you rolled (1€, ..., 6€). Suppose the result was one. What would you do? Unless you suffer from intrusive feelings of being observed even when you are not, or are fine with your social status as an antisocial person, you would probably report rolling four or five [79]. Why not six? The reason for that is such a report would be risky. While it might be the case you actually rolled six, you would nevertheless be a blatant liar with probability $\frac{5}{6}$. By avoiding this your conscience serves you its role — maximizing profit without risking too much (for a review of cheating research see also [80]).

These are strange and often funny results, but this is what there is to investigate in the study of moral cognition. Observed patterns of behavior reveal underlying tradeoffs between willingness to violate the interests of others and defending oneself against moralistic aggression. They also suggest that consciously realized moral rules have little or nothing to do with actual behavioral predispositions. Who thinks and admits that publicly that cheating and lying is permissible? Moral rules are only superficial representations and their utterance is supposed to signal one's trustworthiness and good value as a potential social partner. The rules people report to believe in and the rules people live by need not be the same. This discrepancy between attitudes and behaviors poses a serious problem for moral psychology [81]–[83].

Most importantly, this internal motivational conflict points out that in the study of moral cognition, the content of moral rules does not tell the full story. Instead, it serves as a function of

adaptations that cause humans adopting it (see also [84]). The questions that research in moral psychology needs to answer is: (1) what proximate mechanisms guide decisions in the context of conflicts of interests, (2) what is the ultimate cause of a particular disposition, and (3) how it relates to people private beliefs regarding moral issues. From this perspective, it needs to be emphasized that research in moral psychology does not investigate moral norms in the first place. Rather, instead of treating moral norms as explanations, they should be understood as phenomena to be explained.

These considerations brings us to an important distinction between *moral behavior* and *moral judgment*. While the first concerns the question what people *would* do in a particular situation, the second concerns the question what people think they *should* do (see also [50]). The study of judgment is far more tricky - while the research in moral psychology has moved beyond superficial understanding of morality as moral norms and moral reasoning, focusing on the proximate emotional mechanisms producing judgments [85], [86], the task of investigating causes of it is not trivial. Morality is a weapon that can turn against its constructors, and the content of moral beliefs and judgment reflect this fact. Its generation relies upon simulating yourself in a situations where you want to constrain others (*aggressive morality*) and situations where you want to justify your own actions (*defensive morality*).

We cannot say that physically attacking others is always permissible because we would like to have some norms against people who attack *us*. On the other hand, we also cannot say that attacking others is always wrong because sometimes we want to perform this action toward *others*. This conflict leads to numerous complex conditional rules that reflect negotiation between these two motivations (e.g. murder-manslaughter distinction, self-defense killing, capital punishment). Figure 2.2 depicts the process of aggressive use, defensive use, and impartial judgment generation.

The process of generating moral judgments indirectly but ultimately depends on agent's interests concerning the situation outcome. There is variability in how people judge particular cases and this variability is related to the roles they are associated with. Some of them will consider themselves more likely to be in the role of interest violator, some of them will likely become victims. These different interests produce different moral norms.

Parents often think that their children *should* clean their rooms, while the children disagree. Copyright owners think it is *immoral* to steal their content on the Internet, users do not see this problem. Terrorists think it is *justified* to kill civil people if the cause is worth it, families of their victims disagree. Investigators think it is *permissible* to use *enhanced interrogation techniques*, criminals and their associates think it is not. Of course, all of these people will explain their

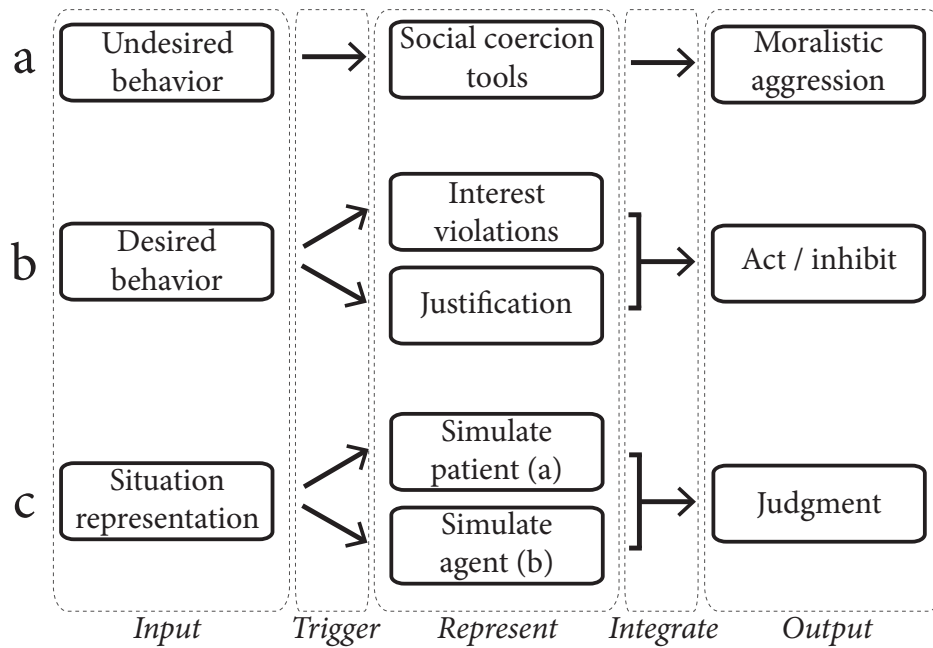


Figure 2.2: The Uses of Moral Faculty: Panel a represents the aggressive use of moral faculty performed by moral patient whose interests were violated; panel b represents the defensive use of moral faculty by moral agent who considers performing interest-violating action; panel c represents the integrative use if moral faculty in which judgment is produced.

	PHENOMENON	EMOTIONS	REPUTATIONAL CONCERNS
<i>Aggressive</i>	Other-directed reaction: – Moralizing – Accusing	– Contempt – Anger – Disgust	Critique of others in order to imply one’s moral superiority
<i>Defensive</i>	Self-directed control: – Refraining – Concealing – Justifying	– Guilt – Shame – Embarrassment – Fear	Signaling the importance of rules without willingness to conform
<i>Integrative</i>	<i>Impartial</i> judgment	– Empathy – Schadenfreude (...)	

Table 2.3: Components of Moral Faculty: This table summarizes components of the moral faculty discussed in the current Chapter.

behavior in terms of applying moral norms. The goal of this chapter was to illustrate that it would be profoundly wrong to treat these judgments as causes rather than effects of particular behavioral predispositions.

To summarize, the problem with moral judgment interpretation is that: (1) we are biased by our interests, (2) we produce judgments that show us in a positive light without the willingness to apply them to ourselves, and (3) we may incorrectly forecast the importance of proximate emotional mechanisms in real world situations [83].

Painting the full picture requires not only describing superficial content of judgments or disagreements, but also identifying *ultimate* and *proximate* mechanisms producing *moral behavior* and influences on judgment integration processes. In Table 2.3 I summarize functions of the moral component discussed in this Chapter and indicate what reputational concerns may bias results of studies aiming to investigate these. In Chapter 3 I will review the strategies employed to answer these questions.

Chapter 3

Investigating Morality

In the present chapter we will discuss main approaches to conducting research on the moral faculty. We will review prevalent methods and ask the following questions: (1) what is research method *supposed* to measure?, and (2) what does the research method actually measure?

Moral cognition is for moral action. It is important to note at the beginning that *moral behavior* and *moral judgment* are related but distinct phenomena. As we have shown in Chapter 2, a discrepancy between human action and its evaluation should be expected. For that reason, it is neither possible to infer from judgments people make how they will actually behave, nor it is possible to infer judgments of right and wrong from human behavior. However, these questions should be seen as complementary, as they can influence each other (see Fig. 2.2).

As it should be clear at this moment, one particular approach to investigate moral cognition can be discarded immediately: Due to the conflict between what people do and what people want others to think they do, one cannot rely on self-reported moral behavior (see [87], [88]). As is the case with every personally sensitive question, such reports will be unreliable due to either conscious lying or unconscious self-deception and positive illusions [19], [76].

If one still asks such questions, the only information one can obtain is what people think is the social norm expected from them. Almost everyone thinks that it is good to keep promises, care for one's family, and protect the environment, and that it is bad to drive when drunk, cheat on your significant other, and accept bribes. The problem is that this tells nothing about an individual's willingness to behave in these ways in real life, as the opportunity or need to make a choice arises.

In topical research, one therefore needs to employ more indirect methods in which the studied behavior will be more natural and judgment less affected by expectancy effects. In the next sections we introduce methods used in moral psychology. In Section 3.1 we describe and evaluate vignette-based methods that aim to study impartial moral judgments. In Section 3.2 we

describe the game-theoretic approach to study moral decisions within the context of economic games. Finally, in the Section 3.3 we introduce the study of moral decisions with Virtual Reality technology.

3.1 Vignettes

One possible approach to investigate moral cognition is to present people with a vignette describing a moral dilemma and see what they say about it. It has a clear advantage over self-reports because it does not pose a direct threat to participants' self-perceptions. A dilemma presented is a hypothetical situation in which people usually need to make a binary choice. In such a way, by carefully construing a conflict within a dilemma, researchers can see what moral concerns take precedence in human judgments.

This research approach has become popular in moral psychology since the publication of an influential paper by Greene, Sommerville, Nystrom, *et al.* [89]. There exist several standardized dilemmas that have been employed in research numerous times in recent years [90], [91]. Due to their artificiality and hypothetical nature, they are considered a valuable approach capable of distilling complex moral distinctions in a compact way.

The most popular and widely studied dilemma in recent years is *The Trolley Dilemma* [92]. The following passage presents the standard variation of it:

You are at the wheel of a runaway trolley quickly approaching a fork in the tracks. On the tracks extending to the left is a group of five railway workmen. On the tracks extending to the right is a single railway workman.

If you do nothing the trolley will proceed to the left, causing the deaths of the five workmen. The only way to avoid the deaths of these workmen is to hit a switch on your dashboard that will cause the trolley to proceed to the right, causing the death of the single workman.

Is it appropriate for you to hit the switch in order to avoid the deaths of the five workmen?¹

This dilemma was supposed to test people intuitions concerning a deontological and utilitarian approach to normative morality. The study investigated the importance of emotional processing in moral judgment contrasting this (so-called *impersonal*) dilemma was contrasted with another (*personal*) variation, in which the person to be sacrificed was a fat man standing on the footbridge, who could be physically pushed to stop the trolley with his body (*The Footbridge Dilemma* [89]). The main finding of the study was that people were less willing to sacrifice one person in order to save five when the dilemma itself evoked stronger emotional response.

¹Available as an online supplementary material to [89]: <http://science.sciencemag.org/content/293/5537/2105.figures-only>, 2016/06/17

But several things remained unclear. It was not really known what exactly caused the difference in people's responses to the *Trolley*, as opposed to the *Footbridge* case. In addition to different emotional reactions caused by the thoughts of pushing a living being and pressing a switch, these dilemmas differed importantly in regard to their causal structure.

Specifically, in the standard *Trolley* case, an actor who decides to kill one worker might say that this outcome was not intended. The aim of the chosen action rather having been to divert the trolley and this action would not have killed anyone if that one worker had not been in the wrong place, at the wrong time. In such a way his death is only an unintended side-effect. Unfortunately, this justification does not apply to the *Footbridge* case. There, one cannot achieve much without the fat man. If he was been there, one would not have a real choice. Here the act of killing therefore is intentional, the fat man being used as a means.

In order to clarify this matter, more variations of the dilemma were needed. One clever way to address this problem was *The Loop Track* variation [93, p. 149], see also [94]:

Ned is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. The heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die. Is it morally permissible for Ned to throw the switch?

This dilemma is fundamentally different from the standard *Trolley* case. Here, the diverted train will return on the main track if not stopped by anything. Because of that, killing the one worker can no longer be justified as a side-effect. According to the study, for some people it makes a difference.

But what about physical contact? It might be the case that for people this is a morally significant distinction. Comparison of the *Trolley* and the *Footbridge* cases is not sound since physical contact is present only in one. More dilemmas were needed. This time, problem the issue was addressed by means of a door trap right under the fat man's feet [95, p. 1083]:

Is it permissible for Evan to pull a lever that drops a man off a footbridge and in front of a moving boxcar in order to cause the man to fall and be hit by the boxcar, thereby slowing it and saving five people ahead on the tracks?

There also exists another variation of the door trap case, in which the footbridge with the man standing on a door trap is over the side track (in which case death through the door trap is a side-effect) [95]. Finally, the impact of the use of physical force itself is not clear. Different variations of the *Footbridge* case differentiate between situations in which you push the fat man

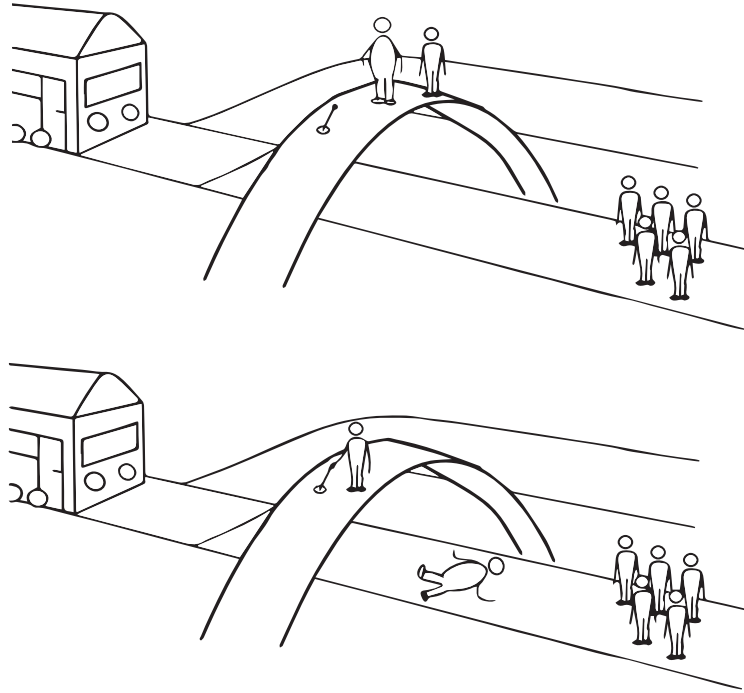


Figure 3.1: Example Variation of the Trolley Dilemma: There is a switch working just like in the standard *Trolley* case but it is located on the footbridge. An actor needs to run to it and accidentally cause the fat man to fall. Dilemma used in [96].

with your bare hands or using a pole, between cases in which the button opening the door trap is right next to or far away from the person to be killed, and between cases where the switch is located on the footbridge and running to it you need to intentionally or accidentally run into the fat man, causing him to fall and die [96] (see Fig. 3.1).

All of these dilemmas have non-trivial causal structures that can provoke a lot of interesting moral reasoning. The examples given here are by no means exhaustive — many more scenarios can be generated from the base case. A lot of work in moral psychology has been invested in structural modeling and detailed discussion of situation representations for such dilemmas [93], [97], [98].

But the key issue is following: as researchers have ended up studying people’s responses to opening a door trap under someone’s feet or accidentally pushing a fat man in order to turn the switch located on the footbridge, what do these research methods *actually* measure?

It seems that after moving from realistic, real-world experiences to the world of fiction, the psychological mechanisms evoked by these dilemmas change. The main reason for this fact is that people do not take these dilemmas seriously. Even in case of the standard dilemma (no door traps) the proportion of people who deny various aspects of it or devise alternative solutions is

not negligible [99]. As people read the vignettes, they do not think about moral concerns within, they are thinking that the things described there just cannot happen in the real life. When they respond, instead of making a serious moral judgment, they make fun out of it — in one study participants were asked how many people would have to be on the tracks in the *Footbridge* case for pushing the fat man to be permissible and more than 10% answered: 0 [100].

These dilemmas may be useful for distilling and illustrating some moral distinctions in a hypothetical and uncontroversial context, but they have little to do with real-life moral judgments. The fact that they are unrealistic and unrepresentative for everyday moral decisions may be an issue, but the fact that they do not engage the psychological processes involved in ordinary moral judgment renders them useless for purposes other than (moral-philosophical) entertainment (see [101]).

Moreover, even if these dilemmas had some connection to the reality we live in, what would responses measure?

An additional problem with giving responses to such enquiries, even in hypothetical contexts, is that the method also suffer from *expectancy effect*. It is not really known what influences these judgments are exposed to. For instance, it was found that results can be affected by the way question is asked (“would you” *versus* “is it permissible”) [102]. They can also be affected by the specified actor (“what would *you* do” *versus* “what *someone* should do”) [100], [103]. Finally, there might be non-utilitarian reasons for giving utilitarian responses in these dilemmas that have not been taken into consideration [104], [105].

Researchers have done a an interesting job formalizing the structure of moral dilemmas and legal or moral concepts underlying them. Some of these distinctions clearly have an impact on *some* people’s decisions [93]. But in the case of ordinary people, it is not known how many of them really understand what they are meant to be asked about. Research on hypothetical dilemmas tests how people rationalize (or not) intuitive moral judgments [95]. But as these peculiar judgments are not related to real-life experience, they tell us nothing about applying moral distinctions in social context: When and how do we use unintended side-effect justification? When is it successful in defending one’s reputation?

The main advantage of *trolleyology* is that these questions do not pose a threat to the participants’ self-image. If people were presented with a familiar dilemma involving real moral issues (*should a policeman accept a bribe?*), the answers would be predictable and useless: They would answer in a socially desirable way. The fact there is no socially desirable way to answer *The Trolley Dilemma* makes it a potentially useful method. Unfortunately, because its narrative is more applicable to a cartoon than a real-life story, they do not provoke any serious judgment.

As Bauman, McGraw, Bartels, *et al.* [101] note, serious moral dilemmas provoke strong emotions and moral outrage among people who disagree. But if none of us will ever be in the situation described in *The Trolley Dilemma*, what is there to fight about?

Naturally, not all of studies employing text-based methods rely on scenarios as unrealistic as *The Trolley Problem*. One can legitimately ask at this point whether the failure of the method in studies described so far should be attributed to the inherent weakness of the method or merely to the particular choice of scenarios.

While we agree with the contention that trolley-like scenarios do not provide any interesting insight about human morality (see [101]), we think that the problem is even more fundamental and is caused by the very nature of text-based methods. This is the case due to the reasons discussed in Chapter 2 — if morality is a tool employed in dynamic interactions among people, it should be scientifically investigated in these interactions as well.

3.2 Behavioral Game Theory

One important feature of real-world judgment and decision-making was absent in vignette-based studies: The person making the judgment had no personal interest in one outcome over another. As illustrated in Chapter 2, this is a defining feature of making both behavioral and hypothetical decisions. A study of moral cognition needs to investigate invention and the (motivated) use of principles not in the abstract, but in real-world settings. Even if you are an observer, you are not completely unaffected by the conflict resolution. If people were put in a situation defined by a particular conflict of interest, their responses would become more realistic.

A study conducted by FeldmanHall, Mobbs, Evans, *et al.* [106] illustrates the point. The setup of the study was the following: Participants were given 20 £ and informed that another person would receive painful electric shocks. Then they could make a decision of contributing some or all of their money to alleviate or completely remove the strength of the electric shocks. The result was that, on average, subjects paid 7.48 £. In another condition, subjects were not actually paid but only asked hypothetically what they would do in the described situation. As could be expected, the average claim was that they would contribute 18.47 £.

That is a *huge* difference. The reason for it was that in the case of the hypothetical scenario, the only motivation behind giving an answer was social expectation to treat people's well-being as higher value than money. Knowingly to themselves or not, they reported something divergent from what they would really do. Hypothetical cases cannot evoke the same proximate emotional mechanisms and thus people's responses are controlled differently [107]. This also illustrates the argument we put forward in Chapter 2 — people may believe that human life is more important

	C	D
C	a, a	c, b
D	b, c	d, d

Table 3.1: The Prisoner’s Dilemma: $b > a > d > c, 2a > b$. C stands for cooperate, D stands for defect. It is most beneficial for an agent to be exploitative towards others while they constrain themselves, but if everyone starts to behave that way, it leads to a suboptimal aggregate outcome. Adapted from [109].

than their own selfish interests, but in reality this belief may itself be based on moral reputation building.

If we want to investigate moral decisions, there must be real stakes behind these decisions. Realization of this fundamental requirement underlies research in behavioral game theory [108]. The framework of games highlights a very important feature of moral situations — people are interdependent. As we argued before, related uncertainty is inherent to making decisions in the real world. Can I trust my partner? Are there any cameras or witnesses? Will I be caught if I ride the metro without a ticket? In the real world, you cannot always know answers to these questions. What we do depends on what we think others will do.

Out of the numerous games used to investigate moral decision-making, arguably the most prominent one is *The Prisoner’s Dilemma*, and it will serve as an example here (see [51]). The payoff matrix defining the game is presented in Table 3.1. Its original formulation was based on a scenario in which two criminals are captured and kept separately. Then, each of them is given a choice: he can snitch on his associate (*defect*) or remain silent (*cooperate*). If he defects while his partner does not, he is set free while his associate goes to jail for 10 years. If he cooperates while his partner also cooperates, both of them goes to jail for 2 years. If he defects while the other also defects, both of them go to jail for 6 years. But if he cooperates while his partner defects, the partner is set free and he is sent to jail for 10 years.

The beauty of this game lies in the fact that it captures a complex logic of human conflict in a simple formulation: *I* know it is rational to defect, since regardless what *You* do, *I* will always be better off doing that. But *I* also know that *You* probably realize the same. Do *You* know that *I* know that *You* know it is rational to defect? Considerations inherent in this game lead to an endless recursion of mental state attributions.

How can it happen that at least one of us defects? It might be that *I* want to decrease my sentence. Immediate self-interest is an important motivator that can easily override moral

obligations not to violate interests of others. In the end, if circumstances were tough: would *I* be really responsible?

It can also happen that *I* do not want to violate Your interests, but *I* think *You* will violate mine and defect. In such a case, *I* have no other choice as to perform a (*morally justified*) defection in order to protect myself against the worst outcome.

Finally, it might also happen that *I* do not care so much about my own well-being but *I* simply do not like *You*, think that *You* deserve the worst, and cannot resist the pleasure of knowing *You* suffer. In that case, *I* will also perform a (*morally justified*) defection.

A superficial look at *The Prisoner's Dilemma* may suggest that one needs to be selfish in order to defect. But there are also other motives leading to that outcome, creating an ambiguous situation (see [110]). These motives illustrate the defensive side of morality discussed in Section 2.2. Once one has identified a most beneficial outcome, one will always find a justification supporting it. That is the power of defensive moral reasoning.

A generalized version of *The Prisoner's Dilemma* is *The Public Goods Game* [111]. In this game, a group of players is initially endowed with some sum of money they can invest into a common pool. After everyone made an investment, the amount of resources in the common pool is increased such that the marginal per capita return (*MCPR*) is smaller than 1² and distributed equally among all players. In such a way, group's collective welfare is maximized when all of members invest as much as possible, whereas the best option for an individual is to invest nothing while others invest everything. This game represents a famous *Tragedy of the Commons* situation [112] and triggers strategic thinking patterns analogous to *The Prisoner's Dilemma*.

The study of decision-making within a game seems to be then a good approach to capture the conflict between willingness to perform a justified violation of others' interests and fear of being attacked by moralizers (see Fig. 2.2, panel b). These attitudes can be directly derived from the payoff structure. Due to its abstract formalization, this setup also ensures that everyone agrees about the desirabilities of particular outcomes: as they are quantified, they can be easily compared and there is no place for an ambiguous interpretation. Moreover, a game may comprise multiple stages, such as a punishment stage, which adds interesting complexity. Finally, participants' earnings *actually* depend on their choices, so differently from vignette-based studies, there do exist a motivation other than tweaking one's judgment to promote one's own social reputation.

These properties would appear to make game-theoretic modeling the candidate of choice

²This is to ensure contribution is not a payoff-maximizing strategy.

for the study of strategic interaction and moral concerns. Indeed, the wide applicability and conceptual possibilities of these games have led to the popularity of this method [for a review, see 113].

Unfortunately, this approach also has some important shortcomings. These are due to a prosaic fact — people often lack the working competence to solve such formally specified games [114]. As has been illustrated by an investigation for the case of *Public Goods Game*, it is estimated that half of the people cannot identify the payoff-maximizing strategy [111], [115]. In public goods studies that use *M CPR* values higher than 1, thereby removing the entire dilemma and ensuring that contribution is a payoff-maximizing strategy, people still do not contribute their entire endowment [116]. While it is possible they do that out of spite and are willing to decrease their own payoff only in order to harm others [117], [118], it can also happen that they choose some random amount “roughly in the middle” without understanding how the game works. This interpretation is supported by a study which found similar contribution levels when people were informed they would be giving money to a randomly-deciding black box [119]. It has also been found that the level of game rules misunderstanding is correlated with subsequent *prosocial* behavior [120].

Moreover, even for people who understand the game, it is not clear what their decisions mean. Expectancy effects and awareness of being monitored poses a problem for all research involving human subjects, but the study of social behaviors is the most susceptible area [121]. Framing of the game and other contextual factors can exert more influence than the payoff structure itself [113], [121]. It is not known how these behaviors generalize to different contexts. Last but not least, laboratory environments are not typical for human interactions and interpretations of the results of the research need to take this into consideration.

This problem is probably most obvious in the study of one-shot social dilemmas, in which participants are informed they are interacting with their partners anonymously and only once. Due to the fact that humans have evolved in an environment where anonymous interactions are not typical, they have an evolved bias towards treating people as possible future partners [122]. This automatic behavior would need to be consciously overridden. Missing this fact and treating people’s responses to laboratory games as realistic has led to numerous wrong interpretations of experimental data [123], [124].

To summarize, research in behavioral game theory provides methods ensuring real consequences of people’s actions. Unfortunately, due to an unrealistic environment of experiments, people’s responses cannot be generalized. It is also difficult to establish whether people understand the logic of strategic interactions in ways experimenters want them to. Thus, results

of these experiments are confounded by errors, random guessing, and giving socially desirable responses.

3.3 Virtual Reality

As we have shown in Sections 3.1 and 3.2, a central problem affecting research methods investigating moral faculty is that people are placed in unrealistic environments they do not understand. Indeed, researchers have recognized this problem and called for development of better methods (see e.g. [77], [125]). What are the ways to address this problem?

One proposed possibility lies in the employment of Virtual Reality technology as a research method. This approach is currently far less popular than the methods discussed earlier, but it has its promises. The claimed main advantage is that ecological validity and participants' engagement with the research could be higher [126]–[131]. Additionally, this research strategy would enable putting participants in threatening or morally problematic situations that are dynamic [132], [133]. As they are placed in realistic situations that resemble the real world across multiple modalities, people's understanding and engagement might be improved. When virtual worlds are inhabited by virtual characters who are embodied, interacting with them induce a sense of social presence [134].

On the other hand, virtual reality is also qualitatively different from real life, as people are aware their actions will not affect any real person (see [135]). For this reason, empirical studies research how people engage with such settings.

First studies using Virtual Reality bring promising results. Slater, Antley, Davison, *et al.* [136] conducted a study aimed at replicating the social psychology classic *Obedience Experiment* by Milgram [137]. Most importantly, this study reported that people experience negative arousal while performing harmful actions in the virtual environment, even though the 3D models of characters had significant limitations (see also [138]).

The Trolley Dilemma has also been investigated within virtual reality and the results showed interesting deviations from vignette-based findings. Navarrete, McDonald, Mott, *et al.* [139] replicated the standard results but also found that there were participants who first pulled the switch and then, as they observed the resulting situation, changed their minds and returned it to its original position, a situated behavioral pattern that could not possibly occur in vignette-based studies.

Patil, Cogoni, Zangrando, *et al.* [140] found that more people in *The Trolley Dilemma* acted in a utilitarian matter than judged it to be the right thing to do, suggesting higher importance of consequences over obligations in dynamical decision-making situations (see also [141], [142]).

Illustrating the possibility to investigate temporal changes in moral decisions, a study conducted by Zanon, Novembre, Zangrando, *et al.* [143] put participants in a situations where they needed to escape a building during the spread of fire and met a trapped person calling for help. It was found that 20 % of people attempted to save this person but gave up when they realized it to be too risky and time-consuming.

A study by Nee, White, Woolford, *et al.* [144] illustrates the similarity of behavioral patterns within virtual reality and real life. They had experienced burglars and normal people enter the house both in real life and virtual reality. They found similar differences in these two groups in terms of burglary-related behavior. Most importantly, they showed that differences present in real life hold also in virtual reality.

Gabriels, Poels, and Braeckman [145] found that *Second Life* residents feel equally intense self-reported moral emotions in response to being cheated by their partners in virtual reality and real life. Similarly, Cristofari and Guitton [146] studied anecdotal reports on morally problematic actions taken in virtual multi-player games and found that people express guilt and morally justify actions committed there.

In addition, several studies have illustrated the importance for moral considerations on actions taken within virtual reality. For instance, it was found that actions for which players cannot find a good justification produce more guilt [147], [148]. Joeckel, Bowman, and Dogruel [149] modeled moral dilemmas after Haidt and Joseph's [6] moral foundations and found that the salience of different norms can decrease the proportion of violations within a game context (see also [150]).

These results illustrate that people do take actions performed in Virtual Reality seriously, despite the absence of real-world consequences. It happens because realistic response is guaranteed by virtue of what their actions represent [135], [151].

Indeed, some researchers have argued that since games are systems that simulate real-world social environments, they promote the use of moral decision-making and reasoning [152]–[155]. In such a way, this research mainly focuses on developing *serious games* used for training [156]–[158]. In the context of this thesis however, we are interested in the potential use of these games (VR studies) for *research* purposes. These goals are not mutually exclusive, as a game developed for moral decision-making research lends itself also to being employed for discussing implications of their decisions in debriefing, as well as any game developed for training might lend itself to being employed for measuring what choices people *actually* make.

Nevertheless, such approaches to investigating moral faculty also have their problems. A main challenge for researchers employing such a method is to ensure that actions taken within

	<i>Realistic Situation</i>	<i>True Consequences</i>
<i>Vignettes</i>	–	–
<i>Behavioral Game Theory</i>	–	+
<i>Virtual Reality</i>	+	–/+

Table 3.2: Summary of research methods in moral psychology.

game context generalize to the real world. To put this question differently: Are we guaranteed that people would act in the real world as they act within the virtual one?

This is an open research question that is investigated empirically. Even if the environment produced is realistic and people understand the situation better than in the case of the game-theoretical studies described in Section 3.2, it needs to be researched how they engage with the task, as real world consequences are absent: For example, participants can make serious decisions they would make in an analogical situation in the real world or they can engage in *metagaming*, performing exploratory, *what-if* choices [159].

Studies on player engagement in games suggest that there are different behavioral patterns. Weaver and Lewis [160] found that 68% of their study participants claimed they make decisions they would do in real life, while Lange [161] found that only 10% claimed to do it *always* and 55% claimed to do it *usually*.

These results suggest that special care needs to be taken on ensuring that players engage with the game system seriously. There are no definite answers exactly what elements maximize such engagement (see [162]). It seems, however, that the crucial element is the emotional involvement of the player [163]. Players need to interpret these virtual situations seriously; this is best achieved by putting them in realistic and familiar dilemmas. In Section 3.1 we described what scenarios, we believe, not to choose.

To summarize, research on moral faculty conducted within Virtual Reality appears to present a promising avenue due to possibilities of creating realistic and dynamic environments that increase ecological validity of research. Even if the consequences of actions do not extend into the real world, high emotional engagement can create a situation in which people behave *as if* they were in the real world. Such serious treatment of virtual settings by participants is not guaranteed, researchers are working on development of methods that maximize it. The features we have addressed are summarized and compared to other methods in Table 3.2.

In Chapter 4 we introduce a game we prototyped for the purpose of measuring moral decisions and subsequent reasoning patterns.

Chapter 4

The Counterfeiting Game

Having characterized the human moral faculty in Chapter 2 and research approaches employed to investigate it in Chapter 3, in the present Chapter we introduce the prototype of a research instrument developed by us. In Section 4.1 we describe the phenomenon of counterfeiting. We characterize the nature of decisions made in the context of this crime and argue that scenarios modeling it may be a promising avenue. In Section 4.2 we describe our research instrument and in Section 4.3 critically evaluate it.

4.1 Problem Introduction

The scenario that will be implemented will involve the production of counterfeit goods. The choice of this phenomenon as a background for making moral decisions is motivated by several factors [164], [165]:

- Counterfeiting represent a real-life example of prevalent unethical behavior,
- Ethical consequences are often non-obvious due to an indirect violation of interests,
- Legal enforcement is very weak,
- Due to complexity of the supply chain, there are many parties involved in the full production pipeline.

Due to these circumstances, the emergence of unethical decisions regarding involvement in this business becomes very likely. Lack of effective enforcement and relatively small costs related to this business activity motivate many parties, including organized crime groups, to get involved. Moreover, in addition to the violation of property rights and unfair competition, counterfeiting often leads to different secondary undesirable effects, such as sweatshop labor practices

or corruption. All of these may happen with different degrees of awareness in involved actors. Ineffectiveness of legal enforcement ensures that decisions made by actors are not motivated by conscious fear of punishment only.

The business is growing and it has captured an interest from researchers. While the majority of academic research has been concerned with customer decisions regarding the purchase of counterfeit goods [166]–[170], other researchers have begun to identify strategies employed by producers of counterfeited goods [171]–[173].

For the purposes of the current research, we focus on the supply side of counterfeiting. Specifically, our aim is to model social interactions within companies that Staake, Thiesse, and Fleisch [171] classify as *fraudsters* — producers whose aim is to manufacture counterfeit goods of high visual but low functional quality with an intention to deceive their customers. Forged goods of this category include e.g. perfumes or mechanical parts. This type of counterfeiting poses a significant challenge because fake goods can enter legitimate markets through recruiting retailers or using e-shopping [172].

The problem that is faced by a decision-maker corresponds to the *defensive* aspect of morality (see Section 2.2) — an action of counterfeiting is a beneficial one, since it is likely to increase sales of a product, as opposed to using an original but unknown label. At the same time, this action violates interests of copyright owners, turning the problem into a moral one.

4.2 Description of the Game

Employment of the Virtual Reality method for studying moral decision-making opens the possibility of collecting data from responses in dynamic interactions. In other words, while people are immersed into virtual worlds they can be explicitly asked about various aspects of it. These questions might be posed by non-player characters they encounter in the virtual world. As these interactions are put into the context of a story people are placed in, they have good prospect of being interpreted as natural. In such a way, this research method may shed light not only on the specific decision, but also on its antecedents and consequences. Specifically, we believe that in the case of moral decision-making research, any instrument should address three interrelated questions:

- What *mental representations* of a situation do people have?
- What *decisions* do they make?
- How do they *justify* these decisions?

The first item, *mental representation*, concerns questions about people’s mental causal model of the environment. It includes their understanding and awareness of consequences of their actions, as well as expectations of actions of others. In the context of virtual interactions, these representations can be investigated through questions about opinions regarding a situation the player finds themselves in or about their expectation what others would do. The importance of this construct follows directly from the definition of moral issues we introduced in Chapter 2 — as moral concerns arise in situations of *conflict of interests*, any method needs to investigate how people perceive this *conflict of interests*.

The second item, *decision*, concerns obtaining data about one’s actual choice. How can such choice be made in a virtual world? The topic of designing engaging choice systems has gained an attention of commercial game designers and critics [174], [175]. There are numerous possibilities of how such decision mechanism can be implemented — for example, players can be asked to make a choice within a conversation with a virtual character or they can make it by interacting or refraining to interact with a particular object in an environment, or they can make a decision by going to a particular place. The goal of such implementation is to provide a mechanism that feels natural and is not ostentatious — any mechanism that fails to achieve that by unnecessarily broadcasting a “*Now you are making a moral decision!*” message triggers expectancy effects and biased responses. The situation and decision point should be presented in a natural way, and the question whether people recognize the ethical dimension of their decisions should be treated as a dependent variable.

The third item, *justification*, concerns obtaining information about how people make sense of decisions they have made or plan to make. As discussed in Section 2.2, the process of generating justifications is central to moral decision-making situations. Within a virtual world, these justifications may be reported by the players within conversations with virtual characters, after an actual decision has been made. In such a way, the use of virtual reality enables studying justifications as a dynamic social process. Both *Neutralization* [58] and *Moral Disengagement* [59]–[62] theories conceptualized justifications as triggered by a particular situation and produced in reference to it. Despite this fact, researchers in both communities have treated justification production as individual propensity, measured by self-reports [63], [66], [176]. We believe that the virtual reality method may constitute a valid instrument for investigating justifications right after the decision to be justified was taken.

We therefore believe that studying *mental representations* and *justifications* in addition to mere *decisions* is not only an addition but an essential part of investigating human moral faculty. This follows directly from the definition of moral issues we introduced in the beginning

of Chapter 2. Morally significant *behaviors* and *judgments* are those which respond to *conflicts of interests* [3]. Hence, beyond the *moral decision* itself the research question comprises to what extent people are aware of (conflicting) interests of involved parties, what they *expect* from them and how this expectation influences their decisions. We will get back to addressing the question whether the developed game properly measures these in Section 4.3.

With these guidelines in mind, we developed a scenario set in a company producing car accessories that considers engagement in production of counterfeit goods (see Section 4.1). The interactive 3D environment was modeled using the Unity Game Engine¹. Within this non-immersive desktop-based virtual reality system, the player can control their character from a first-person perspective (using the standard *WASD*²—plus—mouse control system). Interactions with non-player characters are possible through the use of a simple menu-based dialog system³—during a conversation, whenever the player needs to make a decision, a menu appears at the bottom of the screen and the player chooses one of the displayed options with a mouse click. In our scenario, such responses need to be made under a time pressure so the player experiences a sense of urgency.

The game starts on a street located in an industrial part of a city (see Fig. 4.1). The character is surrounded by dirty buildings and rusty cars. Urban and construction sounds are played in the background in order to create a multi-modal experience. At the beginning, recorded speech is played to the player which introduces the scenario, the player’s role, and the control system. The player is then instructed to go down the street and start a conversation with the first character (for a full script of the game, see Appendix A).

This first conversation in the game serves the purpose of familiarizing the player with interactions with non-player characters and the dialog system. In this part, no relevant questions are posed, but important way-finding information is offered. This conversation forms a likely hub for stronger social embedding of the player into the story-world. Within this conversation, the player is instructed to enter a nearby courtyard where the car accessories company is located, find their boss inside, and start a conversation with him. When they arrive there, the company’s building is clearly visible. It is surrounded by old cars that are parked, garbage, and some old tires (Fig. 4.2).

The players then enters the building and starts a conversation with their boss. In this conversation, relevant questions are asked and player’s responses are recorded (see Fig. 4.3). The boss remarks the company is not doing well and urges that some action needs to be taken

¹Unity Technologies home page, <https://unity3d.com/>, 2016/06/17

²This refers to these four keys as arranged on QWERTY keyboards.

³Dialogue System for Unity, <https://www.assetstore.unity3d.com/en/#!/content/11672>, 2016/06/17

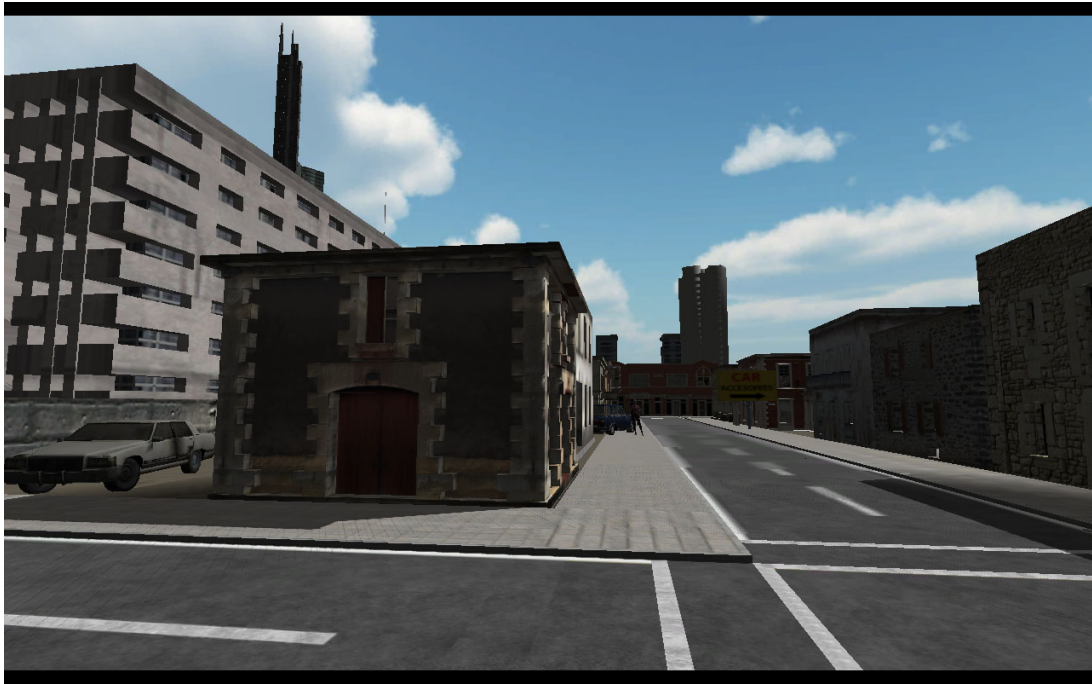


Figure 4.1: Opening Scene of the Game: The player is located in the street and instructed to go forward.



Figure 4.2: The Company Building: The player is instructed to go inside and start the conversation with their boss (who is already waiting behind the entrance).

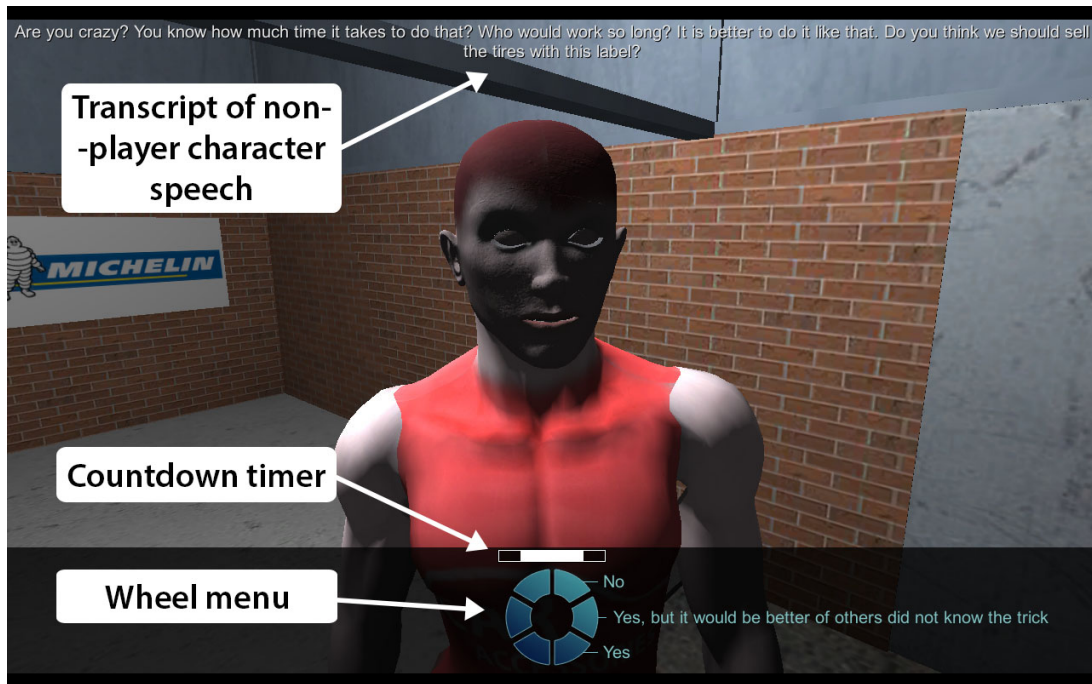


Figure 4.3: Within-Game Conversation: The player chooses their responses from the wheel-style menu that appears at the bottom of the screen; Above it, a countdown timer indicates the time remaining to make a decision (i.e., choose among the options offered); Responses given by non-player characters are displayed at the top of the screen.

in order to change this state of affairs. He presents the player two candidate logos for product re-branding — one that is a novel original, but therefore also without associated brand recognition value, and another one of a widely known company.

The boss suggests that choosing an established label should be beneficial for the company because people associate good quality with it. He advocates it as an easier and popular way to do business. The illegality of such an action is not explicitly mentioned. The player is then asked for their opinion about the practices of the other companies. This question measures the *representation* of the situation.

If the player chooses the answer that most people do not counterfeit, the boss insists on this not being true and that counterfeiting is the best way to make profits, pressuring player into this course of action. If the player immediately subscribes to the view that most parties do counterfeit, the boss then follows up by asking whether they should also sell their tires under the other company’s label. The player is then given a binary choice and their *decision* is recorded.

The next question posed by the boss is meant to measure *justification* and depends on the player’s prior decision: If the player decided not to counterfeit, the boss refers to lost profits and asks for a reason of this decision; alternatively, if the player chose to counterfeit, the boss now

calls the illegality of this action into play and asks what the player thinks about it. The offered responses and their underlying constructs are summarized in Table 4.1.

DID NOT DECIDE TO COUNTERFEIT

<i>Construct</i>	<i>Response</i>
Fear	What if we'll be caught forging the products by the police?
Moral Concern	That's not fair towards the original producer
Reputation	Forging the stuff is not a thing a reputable company would do...

DECIDED TO COUNTERFEIT

<i>Construct</i>	<i>Response</i>
Moral Justification	If we need to move the company forward, effective actions need to be taken!
Euphemistic Labeling / Advantageous Comparison	It's only a label, we are not selling people anything dangerous!
Distortion of Consequences	Nothing bad will happen, we have a good product, so what's the problem?
Displacement of Responsibility	It's <i>You</i> who suggested this action...
Blaming the Victim / Condemnation of the Condemners	Laws are made to protect big players and harm small businesses as ours, I don't care about these!

Table 4.1: Justifications for Moral Actions: Justifications of the player's decision to counterfeit, interpreted according to theories covered in Table 2.1, p. 9.

4.3 Evaluation of the Game

In Section 3.3 we introduced Virtual Reality as a method of psychological research and described its potential benefits. In the current Section, we will critically evaluate whether the game build by us achieves these benefits.

The theoretical advantage of using Virtual Reality as opposed to text-based methods is that it provides a naturalistic environment of decision. This feature is achieved through the *situated* multi-modal experience of observing and moving in a 3D environment, hearing ambient sounds,

and interacting with human-like characters that express their social and affective state through their gestures, posture, proxemics, and voice. The player develops relationships with other game characters and makes time-pressured decisions in reference to them. Such an experience is *qualitatively* different from reading about the existence of other people and interactions with them in written text, and having to rely on imagining how these people and interactions *actually* look and feel like and unfold over time. All of this should contribute to creating an experience of *immersion* in the player [135], [177].

As described in Section 4.2, we tried to realize a prototype implementation — a social interaction scenario in which the decision is embedded was modeled in a 3D-environment populated with other animated virtual characters. During the development work, a number of technical difficulties were encountered.

The biggest problem affecting physical correspondence to the real world concerns imperfections of animations of in-game characters. During development, we ran into the problem of integrating close lip-sync with the dialog system. Because of our tight schedule, we were forced to resign from providing accurate lip-sync for our characters and instead relied on several conversation-related animations that are not specific for any particular message. This problem clearly affects the experience of interaction. For instance, during both conversations in our game it would be useful if non-playable characters could walk and physically refer to things (e.g., the direction in which the player needs to go to meet the boss, the display of the logos). It would also be more realistic if some third person could come in and join the conversation. Unfortunately, such capabilities could not be included due to technical difficulties. This issue is quite serious and has contributed to failures of evoking realistic experiences in past studies (see e.g. [178]).

Other issues involve:

- Imperfect 3D models of the environment and characters;
- Incorrect line-breaking of displayed texts, which currently may run off the screen and is not perfectly synchronized with the played audio;
- Lack of demonstration during the introduction to the game — we currently rely exclusively on a recorded voice to explain everything — adding subtitles and illustrating how dialog menus work would certainly contribute to a better player experience.

After this short coverage of main technical issues encountered in the development of the game prototype that led us to pursue the contingency plan defined for this eventuality, in Chapter 5 we will present a more fundamental methodological reflection on the development and applicability of such instruments.

Chapter 5

Discussion

Within the context of a discussion about developing moral dilemmas in commercial games, Chris Crawford [179] points out that doing this would require meaningfully representing ambiguous relationships between events. As he writes:

Suppose now that you are a game designer wishing to infuse moral elements into your game. To do so, you must show that a moral infraction will ultimately lead to tragic consequences. This, however, requires you to simulate that lengthy and indirect chain of consequences that is so complicated as to be difficult to perceive. How in the hell do you expect to replicate causality of such vast complexity? The level of causal complexity in games is primitive. Most causal relationships in games are essentially boolean in nature; only the simplest and most direct relationships get the full arithmetic treatment. Fantasies of incorporating morality into our games are about as realistic as a worm dreaming of flying to the moon. Someday, maybe — but certainly not in the foreseeable future. [179]

In Section 3.3, we introduced the Virtual Reality method with a main concern regarding serious treatment by players. While this *methodological* issue will need to be addressed once a completed product version of the scenario is available, engaging in the experience of building a prototype has led us to conclude that a main challenge lies in building a good virtual environment in the first place.

By this statement we do not exclusively mean technical difficulties inherent to programming a *physically* realistic environment and non-playing characters. While these are undoubtedly relevant (see Section 4.3), we believe that even more profound problem is posed by the necessity of developing a *psychologically* realistic world.

By this we mean the problem of making the player understand the world, the events that have led to the decision point in the game, and relationships with other game characters and objects. By necessity, getting to know and engaging with the game can be only allowed to take

a limited amount of time. Players are put into a situation where past events are only concisely introduced in the instructions. Characters with whom the player interacts are met for the first time and the player does not have any emotional connection to them. Can such a situation possibly suffice for evoking psychologically realistic responses?

Of course, any alternative method would face similar problems. Every person interacting with a research instrument brings in their own experiences. The aim of a method is certainly not to exclude the relevance and importance of these, because it is supposed to measure judgments and behaviors of that very person. On the other hand, the method cannot leave that person confused or unclear about their role and the impact of decisions they are supposed to make. As we saw in Sections 3.1 and 3.2, this is exactly what happens in case of text-based and game-theoretic studies. Can the virtual reality method be clearly superior in all respects?

Unfortunately, for reasons outlined by Crawford [179], we must concede the answer is no. Even the most graphically realistic environment with remarkable non-player character animations known from modern commercial games will never fully achieve that. The rationale for this claim is that context will always be missing to a significant and unacceptable extent.

What does this mean for the game we built? In addition to modeling the apparent core of the decision environment itself, we would need to present the player with more of the story that led to this particular decision point. We would need to somehow lead the player through getting involved in the car industry, presenting scenes introducing him to the market, letting them face and experience difficulties inherent to leading a business, letting them meet other colleagues and observe their experience. In order to evoke a single point of decision whether to engage in counterfeiting, it may seem that we would need to get players through extended periods of engaging game-play. Creating a tool that can provide that is certainly out of reach for small research groups, especially when researchers involved are not experts in interactive story-world development.

Naturally, these difficulties have been identified in the past and researchers have proposed different solutions. One possibility is to *mod* existing games instead of building a new environment from scratch (see e.g. [180]). This option certainly alleviates a number of issues, as new stories can then be developed and integrated with existing game mechanics, but they face one significant problem — games that are used as a basis for modding are usually either unrealistic and unrepresentative for the every-day life of modern humans (e.g. *The Elder Scrolls* series), or they are typically limited to warfare context only. For these reasons, we think that if one wants to develop virtual scenarios supposed to reproduce real-life experiences, mods of existing games will not be able achieve the required performance in most cases.

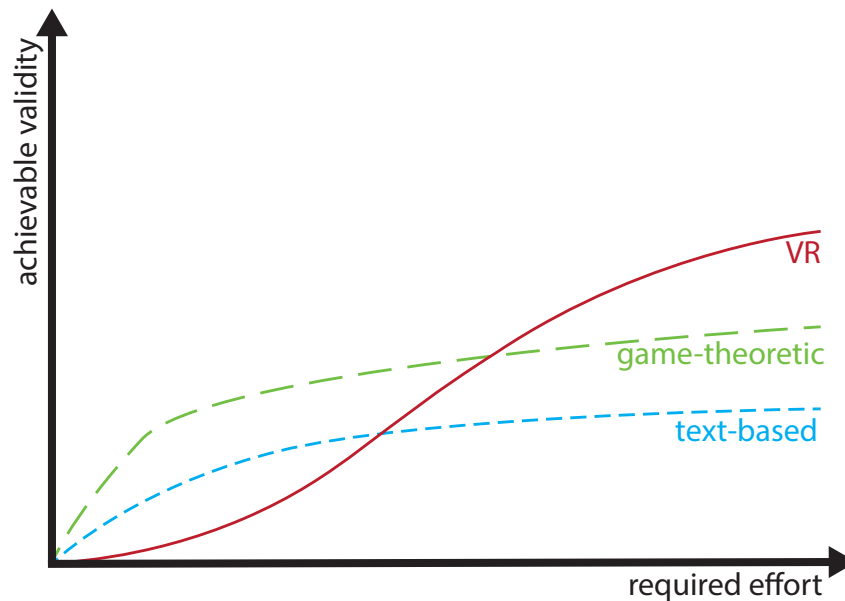


Figure 5.1: Effort-Validity Tradeoff: This plot represents the highest ecological validity of methods yielded by invested required efforts for the different methods addressed in this thesis.

The theoretical advantage of the Virtual Reality method is undoubtedly compelling. Availability of modeling and game development tools, and abundance of materials to learn them with strong community support on the Internet is certainly encouraging for undertaking the task of developing such a method. Such ease of access and the prospect of achieving good results was something that motivated me to start working on this thesis. For sure, this availability of tools makes it possible for any motivated person to assemble their working game. Unfortunately, the amount of work necessary to make the environment and game mechanics so compelling as to outcompete any other method turns out to be so huge that it provokes a question whether it is worthwhile to undertake.

Figure 5.1 represents our estimate of the effort-validity tradeoff in method development, based on our experience in building the virtual reality method. Text-based methods start with some level of validity after small effort of writing a first version of the vignette. Subsequent refinements typically do not significantly change it.

In the case of Virtual Reality, some effort needs to be invested in choosing appropriate tools and packages, and getting them up and running. A lot of effort (compared to producing a vignette) needs to be invested to become first acquainted with and then productive in the use of the tools, so as to produce several minutes gameplay like the one described in the present thesis. Such an approach, due to reasons indicated in the preceding paragraphs, is likely destined to yield similarly poor level of method validity. In order to materialize the potential advantages of the Virtual Reality method, one would need to invest time in developing huge worlds in which

long and rich stories are placed. If one wanted to reproduce the world and social interactions in high detail, this would quickly lead to an effort comparable to the development of modern super-productions like the *Grand Theft Auto* series, *The Witcher* series, or *Fallout* series.

Single-scene virtual reality studies, as described in Section 3.3, as well as the environment built by us, fail to exploit the potential advantages of the method. We believe that these can be achieved by building a coherent story that leads to a given decision point. Studies that attempt to introduce shortcuts based on introducing the preceding plot in the game introduction or game briefing, fall into the same trap of depriving the entire study of its context, much as the vignette-based and game-theoretic methods do. Providing such a context must not be limited to the single scene where the decision is made, it needs to be based on the full story implemented in the virtual world.

Only after the player was engaged in playing such a game, it might be reasonable to investigate whether they understood the environment (*mental representation*), whether their *decision* was taken seriously, or whether their *justification* was based on interaction between their beliefs and the environment.

As any moral concern involves other people, the nature of these people needs to be made clear. Having this requirement in mind, one needs to appreciate the ease with which text-based methods can (to some extent) achieve that — in one paragraph of a vignette one can introduce a coherent story that makes people understand what is their supposed relation to the story characters. In the case of our game, such information might include facts that the boss exerting pressure helped the player in the past, that he is a good and respectable person, etc.

Naturally, it would be an ecologically superior alternative to, instead simply writing these as given facts, let the player experience all of it within the virtual reality. But when one discovers how extensive the story needs to be in order to create one decision point, the virtual reality method ceases to be attractive.

We hope that the main contribution of this thesis will be to serve as an *accessible* caution for researchers considering the virtual reality method. One must not fail to consider how much story *around* the decision itself needs to be developed to make the entire situation *psychologically* realistic. Only if this is implemented and the player starts to perceive and reflect on relations between within-system events, does the validity of the method actually increase. As long as the system is based on modeling single scenes only, the gain from using virtual reality will be only marginal (see Fig. 5.1).

The theoretical advantage of the method remains beyond any reasonable doubt to us. And it is certainly applicable to situations where a single decision is to be made. Modeling of physics-

based challenges, as in learning to drive a car or to fly an airplane is something that can be achieved. But once we enter the social world where the context and the personal history of interactions is relevant, the development of such a system may no longer be viable, in particular when one cannot reliably and substantially draw on the individual background of users.

The motivation and results of virtual reality studies may be fascinating and convincing, but unless these studies address the fundamental issues discussed here, the entire exercise will merely consist of making the best of a bad job.

Bibliography

- [1] F. Nietzsche, *Beyond Good and Evil: Prelude to a Philosophy of the Future*, R. Horstmann and J. Norman, Eds., ser. Cambridge Texts in the History of Philosophy. New York, NY: Cambridge University Press, 1886/2002.
- [2] R. Wright, *The Moral Animal: The New Science of Evolutionary Psychology*. New York, NY: Vintage Books, 1995.
- [3] R. D. Alexander, *The biology of moral systems*. Hawthorne, NY: Aldine de Gruyter, 1987.
- [4] B. Gert and J. Gert, “The definition of morality,” in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Spring 2016, 2016, [Online]. Available: <http://plato.stanford.edu/archives/spr2016/entries/morality-definition/>.
- [5] R. A. Schweder, N. C. Much, M. Mahapatra, and L. Park, “The ”big three” of morality (autonomy, community, divinity) and the ”big three” explanations of suffering,” in *Morality and health*, A. Brandt and P. Rozin, Eds., New York, NY: Routledge, 1997, pp. 119–169.
- [6] J. Haidt and C. Joseph, “Intuitive ethics: How innately prepared intuitions generate culturally variable virtues,” *Daedalus*, vol. 133, no. 4, pp. 55–66, 2004, DOI: [10.1162/0011526042365555](https://doi.org/10.1162/0011526042365555).
- [7] J. Haidt and J. Graham, “When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize,” *Social Justice Research*, vol. 20, no. 1, pp. 98–116, 2007, DOI: [10.1007/s11211-007-0034-z](https://doi.org/10.1007/s11211-007-0034-z).
- [8] P. Rozin, “The process of moralization,” *Psychological Science*, vol. 10, no. 3, pp. 218–221, 1999, DOI: [10.1111/1467-9280.00139](https://doi.org/10.1111/1467-9280.00139).
- [9] J. J. Prinz, “Is morality innate?” In *The Evolution of Morality: Adaptations and Innateness*, ser. Moral Psychology, W. Sinnott-Armstrong, Ed., vol. 1, The MIT Press, 2008, pp. 367–406.

- [10] S. Dwyer, “How not to argue that morality isn’t innate: Comments on Prinz,” in *The Evolution of Morality: Adaptations and Innateness*, ser. Moral Psychology, W. Sinnott-Armstrong, Ed., vol. 1, The MIT Press, 2008, pp. 407–418.
- [11] S. Guglielmo, A. E. Monroe, and B. F. Malle, “At the heart of morality lies folk psychology,” *Inquiry*, vol. 52, no. 5, pp. 449–466, 2009, DOI: [10.1080/00201740903302600](https://doi.org/10.1080/00201740903302600).
- [12] L. Young and A. Waytz, “Mind attribution is for morality,” in *Understanding Other Minds: Perspectives from developmental social neuroscience*, S. Baron-Cohen, H. Tager-Flusberg, and M. Lombardo, Eds., Oxford University Press, 2013, pp. 93–103.
- [13] C. Boehm, *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York, NY: Basic Books, 2012.
- [14] J. Tooby and L. Cosmides, “The psychological foundations of culture,” in *The Adapted Mind: Evolutionary psychology and the generation of culture*, J. Barkow, L. Cosmides, and J. Tooby, Eds., Oxford University Press, 1992, pp. 19–136.
- [15] L. Cosmides and J. Tooby, “Can a general deontic logic capture the facts of human moral reasoning? how the mind interprets social exchange rules and detects cheaters,” in *The Evolution of Morality: Adaptations and Innateness*, ser. Moral Psychology, W. Sinnott-Armstrong, Ed., vol. 1, The MIT Press, 2008, pp. 53–120.
- [16] G. Gigerenzer, “Moral satisficing: Rethinking moral behavior as bounded rationality,” *Topics in Cognitive Science*, vol. 2, no. 3, pp. 528–554, 2010, DOI: [10.1111/j.1756-8765.2010.01094.x](https://doi.org/10.1111/j.1756-8765.2010.01094.x).
- [17] T. C. Scott-Phillips, T. E. Dickins, and S. A. West, “Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences,” *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 38–47, 2011, DOI: [10.1177/1745691610393528](https://doi.org/10.1177/1745691610393528).
- [18] R. E. Nisbett and T. D. Wilson, “Telling more than we can know: Verbal reports on mental processes,” *Psychological Review*, vol. 84, no. 3, pp. 231–259, Mar. 1977, DOI: [10.1037/0033-295x.84.3.231](https://doi.org/10.1037/0033-295x.84.3.231).
- [19] W. von Hippel and R. Trivers, “The evolution and psychology of self-deception,” *Behavioral and Brain Sciences*, vol. 34, no. 01, pp. 1–16, 2011, DOI: [10.1017/s0140525x10001354](https://doi.org/10.1017/s0140525x10001354).
- [20] R. Kurzban, *Why Everyone (Else) Is a Hypocrite: Evolution and the Modular Mind*. Princeton, NJ: Princeton University Press, 2010.
- [21] A. Sell, J. Tooby, and L. Cosmides, “Formidability and the logic of human anger,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 35, pp. 15 073–15 078, 2009, DOI: [10.1073/pnas.0904312106](https://doi.org/10.1073/pnas.0904312106).

- [22] D. F. Luckenbill, “Criminal homicide as a situated transaction,” *Social Problems*, vol. 25, no. 2, pp. 176–186, 1977, DOI: [10.2307/800293](https://doi.org/10.2307/800293).
- [23] M. Enquist and O. Leimar, “The evolution of fatal fighting,” *Animal Behaviour*, vol. 39, no. 1, pp. 1–9, 1990, DOI: [10.1016/S0003-3472\(05\)80721-3](https://doi.org/10.1016/S0003-3472(05)80721-3).
- [24] A. W. Delton and T. E. Robertson, “How the mind makes welfare tradeoffs: Evolution, computation, and emotion,” *Current Opinion in Psychology*, vol. 7, pp. 12–16, 2016, DOI: [10.1016/j.copsyc.2015.06.006](https://doi.org/10.1016/j.copsyc.2015.06.006).
- [25] M. B. Petersen, A. Sell, J. Tooby, and L. Cosmides, “To punish or repair? evolutionary psychology and lay intuitions about modern criminal justice,” *Evolution and Human Behavior*, vol. 33, no. 6, pp. 682–695, 2012, DOI: [10.1016/j.evolhumbehav.2012.05.003](https://doi.org/10.1016/j.evolhumbehav.2012.05.003).
- [26] M. M. Krasnow, L. Cosmides, E. J. Pedersen, and J. Tooby, “What are punishment and reputation for?” *PLOS ONE*, vol. 7, no. 9, T. Zalla, Ed., pp. 1–9, 2012, DOI: [10.1371/journal.pone.0045662](https://doi.org/10.1371/journal.pone.0045662).
- [27] M. E. McCullough, R. Kurzban, and B. A. Tabak, “Cognitive systems for revenge and forgiveness,” *Behavioral and Brain Sciences*, vol. 36, no. 01, pp. 1–15, 2013, DOI: [10.1017/s0140525x11002160](https://doi.org/10.1017/s0140525x11002160).
- [28] P. Rozin, L. Lowery, S. Imada, and J. Haidt, “The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity).,” *Journal of Personality and Social Psychology*, vol. 76, no. 4, pp. 574–586, 1999, DOI: [10.1037/0022-3514.76.4.574](https://doi.org/10.1037/0022-3514.76.4.574).
- [29] J. Haidt, “The moral emotions,” in *Handbook of affective sciences*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds., New York, NY: Oxford University Press, 2003, pp. 852–870.
- [30] D. Pietraszewski and A. Shaw, “Not by strength alone,” *Human Nature*, vol. 26, no. 1, pp. 44–72, 2015, ISSN: 1936-4776, DOI: [10.1007/s12110-015-9220-0](https://doi.org/10.1007/s12110-015-9220-0).
- [31] P. DeScioli, M. Massenkoff, A. Shaw, M. B. Petersen, and R. Kurzban, “Equity or equality? moral judgments follow the money,” *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, no. 1797, 2014, DOI: [10.1098/rspb.2014.2112](https://doi.org/10.1098/rspb.2014.2112).
- [32] K. Bocian and B. Wojciszke, “Self-interest bias in moral judgments of others actions,” *Personality and Social Psychology Bulletin*, vol. 40, no. 7, pp. 898–909, 2014, DOI: [10.1177/0146167214529800](https://doi.org/10.1177/0146167214529800).

- [33] ———, “Unawareness of self-interest bias in moral judgments of others’ behavior,” *Polish Psychological Bulletin*, vol. 45, no. 4, pp. 411–417, 2014, DOI: [10.2478/ppb-2014-0050](https://doi.org/10.2478/ppb-2014-0050).
- [34] A. Rustichini and M. C. Villeval, “Moral hypocrisy, power and social preferences,” *Journal of Economic Behavior & Organization*, vol. 107, pp. 10–24, 2014, DOI: [10.1016/j.jebo.2014.08.002](https://doi.org/10.1016/j.jebo.2014.08.002).
- [35] N. Paharia, K. D. Vohs, and R. Deshpandé, “Sweatshop labor is wrong unless the shoes are cute: Cognition can both help and hurt moral motivated reasoning,” *Organizational Behavior and Human Decision Processes*, vol. 121, no. 1, pp. 81–88, 2013, DOI: [10.1016/j.obhdp.2013.01.001](https://doi.org/10.1016/j.obhdp.2013.01.001).
- [36] J. R. Maze, “The concept of attitude,” *Inquiry*, vol. 16, no. 1-4, pp. 168–205, 1973, DOI: [10.1080/00201747308601684](https://doi.org/10.1080/00201747308601684).
- [37] L. Cronk, “Evolutionary theories of morality and the manipulative use of signals,” *Zygon*, vol. 29, no. 1, pp. 81–101, 1994, DOI: [10.1111/j.1467-9744.1994.tb00651.x](https://doi.org/10.1111/j.1467-9744.1994.tb00651.x).
- [38] J. Tooby and L. Cosmides, “Groups in mind: The coalitional roots of war and morality,” in *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, H. Høgh-Olesen, Ed., Palgrave Macmillan, 2010, pp. 191–234.
- [39] M. B. Petersen, “Moralization as protection against exploitation: Do individuals without allies moralize more?” *Evolution and Human Behavior*, vol. 34, no. 2, pp. 78–85, 2013, DOI: [10.1016/j.evolhumbehav.2012.09.006](https://doi.org/10.1016/j.evolhumbehav.2012.09.006).
- [40] S. Pinker, “The moral instinct,” in *Understanding moral sentiments: Darwinian perspectives*, H. Putnam, S. Neiman, and J. P. Schloss, Eds., New Brunswick, NJ: Transaction Publishers, 2014, pp. 59–80.
- [41] D. S. Gordon, J. R. Madden, and S. E. G. Lea, “Both loved and feared: Third party punishers are viewed as formidable and likeable, but these reputational benefits may only be open to dominant individuals,” *PLOS ONE*, vol. 9, no. 10, T. Mappes, Ed., e110045, 2014, DOI: [10.1371/journal.pone.0110045](https://doi.org/10.1371/journal.pone.0110045).
- [42] J. J. Jordan, M. Hoffman, P. Bloom, and D. G. Rand, “Third-party punishment as a costly signal of trustworthiness,” *Nature*, vol. 530, no. 7591, pp. 473–476, 2016, DOI: [10.1038/nature16981](https://doi.org/10.1038/nature16981).
- [43] J. Marczyk, “Moral alliance strategies theory,” *Evolutionary Psychological Science*, vol. 1, no. 2, pp. 77–90, 2015, DOI: [10.1007/s40806-015-0011-y](https://doi.org/10.1007/s40806-015-0011-y).
- [44] P. DeScioli and R. Kurzban, “Mysteries of morality,” *Cognition*, vol. 112, no. 2, pp. 281–299, 2009, DOI: [10.1016/j.cognition.2009.05.008](https://doi.org/10.1016/j.cognition.2009.05.008).

- [45] ———, “A solution to the mysteries of morality,” *Psychological Bulletin*, vol. 139, no. 2, pp. 477–496, 2013, DOI: [10.1037/a0029065](https://doi.org/10.1037/a0029065).
- [46] P. DeScioli, “The side-taking hypothesis for moral judgment,” *Current Opinion in Psychology*, vol. 7, pp. 23–27, 2016, DOI: [10.1016/j.copsy.2015.07.002](https://doi.org/10.1016/j.copsy.2015.07.002).
- [47] P. M. Patrzyk and M. Takáč, “Cooperation via intimidation: An emergent system of mutual threats can maintain social order,” under review.
- [48] C. D. Batson, *What’s Wrong with Morality? A Social-Psychological Perspective*. New York, NY: Oxford University Press, 2016.
- [49] R. Kurzban and P. DeScioli, “Adaptationist punishment in humans,” *Journal of Bioeconomics*, vol. 15, no. 3, pp. 269–279, 2013, DOI: [10.1007/s10818-013-9153-9](https://doi.org/10.1007/s10818-013-9153-9).
- [50] A. E. Tenbrunsel, K. A. Diekmann, K. A. Wade-Benzoni, and M. H. Bazerman, “The ethical mirage: A temporal explanation as to why we are not as ethical as we think we are,” *Research in Organizational Behavior*, vol. 30, pp. 153–173, 2010, DOI: [10.1016/j.riob.2010.08.004](https://doi.org/10.1016/j.riob.2010.08.004).
- [51] S. Pinker, *The better angels of our nature: Why violence has declined*. New York, NY: Penguin, 2011.
- [52] P. M. Patrzyk, “Artificial moral agents: Current approaches and challenges,” in *Proceedings of the MEi:CogSci Conference 2015*, P. Hochenauer, C. Schreiber, E. Zimmermann, and I. Farkaš, Eds., Bratislava, Slovakia: Comenius University, 2015, p. 124.
- [53] K. Asao and D. M. Buss, “The tripartite theory of machiavellian morality: Judgment, influence, and conscience as distinct moral adaptations,” in *The Evolution of Morality*, T. K. Shackelford and R. D. Hansen, Eds., Springer International Publishing, 2016, pp. 3–25, DOI: [10.1007/978-3-319-19671-8_1](https://doi.org/10.1007/978-3-319-19671-8_1).
- [54] Z. Kunda, “The case for motivated reasoning,” *Psychological Bulletin*, vol. 108, no. 3, pp. 480–498, 1990, DOI: [10.1037/0033-2909.108.3.480](https://doi.org/10.1037/0033-2909.108.3.480).
- [55] E. L. Uhlmann, D. A. Pizarro, D. Tannenbaum, and P. H. Ditto, “The motivated use of moral principles,” *Judgment and Decision Making*, vol. 4, no. 6, pp. 476–491, Oct. 2009, [Online]. Available: <http://journal.sjdm.org/9616/jdm9616.pdf>.
- [56] H. Mercier and D. Sperber, “Why do humans reason? arguments for an argumentative theory,” *Behavioral and Brain Sciences*, vol. 34, no. 02, pp. 57–74, 2011, DOI: [10.1017/S0140525X10000968](https://doi.org/10.1017/S0140525X10000968).

- [57] H. Mercier, “What good is moral reasoning?” *Mind & Society*, vol. 10, no. 2, pp. 131–148, 2011, DOI: [10.1007/s11299-011-0085-6](https://doi.org/10.1007/s11299-011-0085-6).
- [58] G. M. Sykes and D. Matza, “Techniques of neutralization: A theory of delinquency,” *American Sociological Review*, vol. 22, no. 6, pp. 664–670, Dec. 1957, [Online]. Available: <http://www.jstor.org/stable/2089195>.
- [59] A. Bandura, *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood Cliffs, NJ: Prentice Hall, 1986.
- [60] —, “Selective activation and disengagement of moral control,” *Journal of Social Issues*, vol. 46, no. 1, pp. 27–46, 1990, DOI: [10.1111/j.1540-4560.1990.tb00270.x](https://doi.org/10.1111/j.1540-4560.1990.tb00270.x).
- [61] A. Bandura, C. Barbaranelli, G. V. Caprara, and C. Pastorelli, “Mechanisms of moral disengagement in the exercise of moral agency,” *Journal of Personality and Social Psychology*, vol. 71, no. 2, pp. 364–374, Aug. 1996, DOI: [10.1037/0022-3514.71.2.364](https://doi.org/10.1037/0022-3514.71.2.364).
- [62] A. Bandura, “Moral disengagement in the perpetration of inhumanities,” *Personality and Social Psychology Review*, vol. 3, no. 3, pp. 193–209, 1999, DOI: [10.1207/s15327957pspr0303_3](https://doi.org/10.1207/s15327957pspr0303_3).
- [63] C. Moore, “Moral disengagement,” *Current Opinion in Psychology*, vol. 6, pp. 199–204, 2015, DOI: [10.1016/j.copsyc.2015.07.018](https://doi.org/10.1016/j.copsyc.2015.07.018).
- [64] S. Shalvi, F. Gino, R. Barkan, and S. Ayal, “Self-serving justifications: Doing wrong and feeling moral,” *Current Directions in Psychological Science*, vol. 24, no. 2, pp. 125–130, 2015, DOI: [10.1177/0963721414553264](https://doi.org/10.1177/0963721414553264).
- [65] R. Barkan, S. Ayal, and D. Ariely, “Ethical dissonance, justifications, and moral behavior,” *Current Opinion in Psychology*, vol. 6, pp. 157–161, 2015, DOI: [10.1016/j.copsyc.2015.08.001](https://doi.org/10.1016/j.copsyc.2015.08.001).
- [66] S. Maruna and H. Copes, “What have we learned from five decades of neutralization research?” *Crime and Justice*, vol. 32, pp. 221–320, 2005, [Online]. Available: <http://www.jstor.org/stable/3488361>.
- [67] L. E. Eckstein Jackson, “Disengaging from moral disengagement: Scant experimental evidence for a popular theory,” PhD thesis, University of Tennessee, 2012, [Online]. Available: http://trace.tennessee.edu/utk_graddiss/1292.
- [68] J. Dana, R. A. Weber, and J. X. Kuang, “Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness,” *Economic Theory*, vol. 33, no. 1, pp. 67–80, 2007, DOI: [10.1007/s00199-006-0153-z](https://doi.org/10.1007/s00199-006-0153-z).

- [69] J. Andreoni and B. D. Bernheim, “Social image and the 50-50 Norm: A theoretical and experimental analysis of audience effects,” *Econometrica*, vol. 77, no. 5, pp. 1607–1636, Sep. 2009, DOI: [10.3982/ecta7384](https://doi.org/10.3982/ecta7384).
- [70] D. Sperber and N. Baumard, “Moral reputation: An evolutionary and cognitive perspective,” *Mind & Language*, vol. 27, no. 5, pp. 495–518, 2012, DOI: [10.1111/mila.12000](https://doi.org/10.1111/mila.12000).
- [71] L. Caviola and N. Faulmüller, “Moral hypocrisy in economic games—how prosocial behavior is shaped by social expectations,” *Frontiers in Psychology*, vol. 5, 2014, DOI: [10.3389/fpsyg.2014.00897](https://doi.org/10.3389/fpsyg.2014.00897).
- [72] D. D. Johnson, D. T. Blumstein, J. H. Fowler, and M. G. Haselton, “The evolution of error: Error management, cognitive constraints, and adaptive decision-making biases,” *Trends in Ecology & Evolution*, vol. 28, no. 8, pp. 474–481, 2013, DOI: [10.1016/j.tree.2013.05.014](https://doi.org/10.1016/j.tree.2013.05.014).
- [73] D. G. Rand, J. D. Greene, and M. A. Nowak, “Spontaneous giving and calculated greed,” *Nature*, vol. 489, no. 7416, pp. 427–430, 2012, DOI: [10.1038/nature11467](https://doi.org/10.1038/nature11467).
- [74] D. G. Rand, G. E. Newman, and O. M. Wurzbacher, “Social context and the dynamics of cooperative choice,” *J. Behav. Dec. Making*, vol. 28, no. 2, pp. 159–166, Apr. 2015, DOI: [10.1002/bdm.1837](https://doi.org/10.1002/bdm.1837).
- [75] R. L. Trivers, “The evolution of reciprocal altruism,” *The Quarterly Review of Biology*, vol. 46, no. 1, pp. 35–57, Mar. 1971, [Online]. Available: <http://www.jstor.org/stable/2822435>.
- [76] M. D. Alicke and C. Sedikides, “Self-enhancement and self-protection: What they are and what they do,” *European Review of Social Psychology*, vol. 20, no. 1, pp. 1–48, 2009, DOI: [10.1080/10463280802613866](https://doi.org/10.1080/10463280802613866).
- [77] B. Wojciszke, M. Parzuchowski, and K. Bocian, “Moral judgments and impressions,” *Current Opinion in Psychology*, vol. 6, pp. 50–54, 2015, DOI: [10.1016/j.copsyc.2015.03.028](https://doi.org/10.1016/j.copsyc.2015.03.028).
- [78] N. Mazar, O. Amir, and D. Ariely, “The dishonesty of honest people: A theory of self-concept maintenance,” *Journal of Marketing Research*, vol. 45, no. 6, pp. 633–644, 2008, DOI: [10.1509/jmkr.45.6.633](https://doi.org/10.1509/jmkr.45.6.633).
- [79] U. Fischbacher and F. Fllmi-Heusi, “Lies in disguise - an experimental study on cheating,” *Journal of the European Economic Association*, vol. 11, no. 3, pp. 525–547, 2013, DOI: [10.1111/jeea.12014](https://doi.org/10.1111/jeea.12014).

- [80] P. M. Patrzyk, "Would you cheat? cheating behavior, human nature, and decision-making," *Student Pulse*, vol. 6, no. 3, 2014, [Online]. Available: <http://www.studentpulse.com/a?id=871>.
- [81] D. L. Krebs, K. Denton, and G. Wark, "The forms and functions of real-life moral decision-making," *Journal of Moral Education*, vol. 26, no. 2, pp. 131–145, 1997, DOI: [10.1080/0305724970260202](https://doi.org/10.1080/0305724970260202).
- [82] B. Monin and A. Merritt, "Moral hypocrisy, moral inconsistency, and the struggle for moral integrity," in *The social psychology of morality: Exploring the causes of good and evil*, M. Mikulincer and P. R. Shaver, Eds., Washington, DC: American Psychological Association, 2012, pp. 167–184, DOI: [10.1037/13091-009](https://doi.org/10.1037/13091-009).
- [83] R. Teper, C.-B. Zhong, and M. Inzlicht, "How emotions shape moral behavior: Some answers (and questions) for the field of moral psychology," *Social and Personality Psychology Compass*, vol. 9, no. 1, pp. 1–14, 2015, DOI: [10.1111/spc3.12154](https://doi.org/10.1111/spc3.12154).
- [84] F. Cushman and L. Young, "The psychology of dilemmas and the philosophy of morality," *Ethic Theory Moral Prac*, vol. 12, no. 1, pp. 9–24, 2009, DOI: [10.1007/s10677-008-9145-3](https://doi.org/10.1007/s10677-008-9145-3).
- [85] J. Haidt, "The emotional dog and its rational tail: A social intuitionist approach to moral judgment.," *Psychological Review*, vol. 108, no. 4, pp. 814–834, 2001, DOI: [10.1037/0033-295x.108.4.814](https://doi.org/10.1037/0033-295x.108.4.814).
- [86] S. Guglielmo, "Moral judgment as information processing: An integrative review," *Frontiers in Psychology*, vol. 6, 2015, DOI: [10.3389/fpsyg.2015.01637](https://doi.org/10.3389/fpsyg.2015.01637).
- [87] R. F. Baumeister, K. D. Vohs, and D. C. Funder, "Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?" *Perspectives on Psychological Science*, vol. 2, no. 4, pp. 396–403, 2007, DOI: [10.1111/j.1745-6916.2007.00051.x](https://doi.org/10.1111/j.1745-6916.2007.00051.x).
- [88] E. A. Giammarco, "The measurement of individual differences in morality," *Personality and Individual Differences*, vol. 88, pp. 26–34, 2016, DOI: [10.1016/j.paid.2015.08.039](https://doi.org/10.1016/j.paid.2015.08.039).
- [89] J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen, "An fmri investigation of emotional engagement in moral judgment," *Science*, vol. 293, no. 5537, pp. 2105–2108, 2001, DOI: [10.1126/science.1062872](https://doi.org/10.1126/science.1062872).
- [90] J. F. Christensen, A. Flexas, M. Calabrese, N. K. Gut, and A. Gomila, "Moral judgment reloaded: A moral dilemma validation study," *Frontiers in Psychology*, vol. 5, 2014, DOI: [10.3389/fpsyg.2014.00607](https://doi.org/10.3389/fpsyg.2014.00607).

- [91] S. Clifford, V. Iyengar, R. Cabeza, and W. Sinnott-Armstrong, “Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory,” *Behavior Research Methods*, vol. 47, no. 4, pp. 1178–1198, 2015, DOI: [10.3758/s13428-014-0551-2](https://doi.org/10.3758/s13428-014-0551-2).
- [92] J. J. Thomson, “The trolley problem,” *The Yale Law Journal*, vol. 94, no. 6, pp. 1395–1415, May 1985, [Online]. Available: <http://www.jstor.org/stable/796133>.
- [93] J. Mikhail, “Universal moral grammar: Theory, evidence and the future,” *Trends in Cognitive Sciences*, vol. 11, no. 4, pp. 143–152, 2007, DOI: [10.1016/j.tics.2006.12.007](https://doi.org/10.1016/j.tics.2006.12.007).
- [94] M. Hauser, F. Cushman, L. Young, R. K.-X. Jin, and J. Mikhail, “A dissociation between moral judgments and justifications,” *Mind & Language*, vol. 22, no. 1, pp. 1–21, 2007, DOI: [10.1111/j.1468-0017.2006.00297.x](https://doi.org/10.1111/j.1468-0017.2006.00297.x).
- [95] F. Cushman, L. Young, and M. Hauser, “The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm,” *Psychological Science*, vol. 17, no. 12, pp. 1082–1089, 2006, DOI: [10.1111/j.1467-9280.2006.01834.x](https://doi.org/10.1111/j.1467-9280.2006.01834.x).
- [96] J. D. Greene, F. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen, “Pushing moral buttons: The interaction between personal force and intention in moral judgment,” *Cognition*, vol. 111, no. 3, pp. 364–371, 2009, DOI: [10.1016/j.cognition.2009.02.001](https://doi.org/10.1016/j.cognition.2009.02.001).
- [97] J. Mikhail, “Moral grammar and intuitive jurisprudence,” in *Psychology of Learning and Motivation*, B. H. Ross, Ed., Elsevier BV, 2009, pp. 27–100, DOI: [10.1016/S0079-7421\(08\)00402-7](https://doi.org/10.1016/S0079-7421(08)00402-7).
- [98] F. Cushman, “Action, outcome, and value: A dual-system framework for morality,” *Personality and Social Psychology Review*, vol. 17, no. 3, pp. 273–292, 2013, DOI: [10.1177/1088868313495594](https://doi.org/10.1177/1088868313495594).
- [99] P. Danielson, “Surprising judgments about robot drivers: Experiments on rising expectations and blaming humans,” *Etikk i praksis - Nordic Journal of Applied Ethics*, vol. 9, no. 1, 2015, DOI: [10.5324/eip.v9i1.1727](https://doi.org/10.5324/eip.v9i1.1727).
- [100] R. Kurzban, P. DeScioli, and D. Fein, “Hamilton vs. kant: Pitting adaptations for altruism against adaptations for moral judgment,” *Evolution and Human Behavior*, vol. 33, no. 4, pp. 323–333, 2012, DOI: [10.1016/j.evolhumbehav.2011.11.002](https://doi.org/10.1016/j.evolhumbehav.2011.11.002).

- [101] C. W. Bauman, A. P. McGraw, D. M. Bartels, and C. Warren, “Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology,” *Social and Personality Psychology Compass*, vol. 8, no. 9, pp. 536–554, 2014, DOI: [10.1111/spc3.12131](https://doi.org/10.1111/spc3.12131).
- [102] S. Tassy, O. Oullier, J. Mancini, and B. Wicker, “Discrepancies between judgment and choice of action in moral dilemmas,” *Frontiers in Psychology*, vol. 4, May 2013, DOI: [10.3389/fpsyg.2013.00250](https://doi.org/10.3389/fpsyg.2013.00250).
- [103] J. Dana and D. M. Cain, “Advice versus choice,” *Current Opinion in Psychology*, vol. 6, pp. 173–176, 2015, DOI: [10.1016/j.copsyg.2015.08.019](https://doi.org/10.1016/j.copsyg.2015.08.019).
- [104] G. Kahane, J. A. Everett, B. D. Earp, M. Farias, and J. Savulescu, “‘utilitarian’ judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good,” *Cognition*, vol. 134, pp. 193–209, 2015, DOI: [10.1016/j.cognition.2014.10.005](https://doi.org/10.1016/j.cognition.2014.10.005).
- [105] G. Kahane, “Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment,” *Social Neuroscience*, vol. 10, no. 5, pp. 551–560, 2015, DOI: [10.1080/17470919.2015.1023400](https://doi.org/10.1080/17470919.2015.1023400).
- [106] O. FeldmanHall, D. Mobbs, D. Evans, L. Hiscox, L. Navrady, and T. Dalgleish, “What we say and what we do: The relationship between real and hypothetical moral choices,” *Cognition*, vol. 123, no. 3, pp. 434–441, 2012, DOI: [10.1016/j.cognition.2012.02.001](https://doi.org/10.1016/j.cognition.2012.02.001).
- [107] O. FeldmanHall, T. Dalgleish, R. Thompson, D. Evans, S. Schweizer, and D. Mobbs, “Differential neural circuitry and self-interest in real vs hypothetical moral decisions,” *Social Cognitive and Affective Neuroscience*, vol. 7, no. 7, pp. 743–751, 2012, DOI: [10.1093/scan/nss069](https://doi.org/10.1093/scan/nss069).
- [108] C. F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.
- [109] A. W. Tucker, “The mathematics of tucker: A sampler,” *The Two-Year College Mathematics Journal*, vol. 14, no. 3, pp. 228–232, 1983, DOI: [10.2307/3027092](https://doi.org/10.2307/3027092).
- [110] N. Halevy, E. Y. Chou, and J. K. Murnighan, “Mind games: The mental representation of conflict,” *Journal of Personality and Social Psychology*, vol. 102, no. 1, pp. 132–148, 2012, DOI: [10.1037/a0025389](https://doi.org/10.1037/a0025389).
- [111] J. Andreoni, “Cooperation in public-goods experiments: Kindness or confusion?” *The American Economic Review*, vol. 85, no. 4, pp. 891–904, Sep. 1995, [Online]. Available: <http://www.jstor.org/stable/2118238>.

- [112] G. Hardin, “The tragedy of the commons,” *Science*, vol. 162, no. 3859, pp. 1243–1248, 1968, DOI: [10.1126/science.162.3859.1243](https://doi.org/10.1126/science.162.3859.1243).
- [113] P. A. Van Lange, J. Joireman, C. D. Parks, and E. Van Dijk, “The psychology of social dilemmas: A review,” *Organizational Behavior and Human Decision Processes*, vol. 120, no. 2, pp. 125–141, 2013, DOI: [10.1016/j.obhdp.2012.11.003](https://doi.org/10.1016/j.obhdp.2012.11.003).
- [114] G. Devetag and M. Warglien, “Playing the wrong game: An experimental analysis of relational complexity and strategic misrepresentation,” *Games and Economic Behavior*, vol. 62, no. 2, pp. 364–382, 2008, DOI: [10.1016/j.geb.2007.05.007](https://doi.org/10.1016/j.geb.2007.05.007).
- [115] D. Houser and R. Kurzban, “Kindness and confusion in public goods experiments,” *The American Economic Review*, vol. 92, no. 4, pp. 1062–1069, Sep. 2002, [Online]. Available: <http://www.jstor.org/stable/3083295>.
- [116] R. Kummerli, M. N. Burton-Chellew, A. Ross-Gillespie, and S. A. West, “Resistance to extreme strategies, rather than prosocial preferences, can explain human cooperation in public goods games,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 22, pp. 10 125–10 130, 2010, DOI: [10.1073/pnas.1000829107](https://doi.org/10.1073/pnas.1000829107).
- [117] M. Sheskin, P. Bloom, and K. Wynn, “Anti-equality: Social comparison in young children,” *Cognition*, vol. 130, no. 2, pp. 152–156, 2014, DOI: [10.1016/j.cognition.2013.10.008](https://doi.org/10.1016/j.cognition.2013.10.008).
- [118] G. Charness, D. Masclet, and M. C. Villeval, “The dark side of competition for status,” *Management Science*, vol. 60, no. 1, pp. 38–55, 2014, DOI: [10.1287/mnsc.2013.1747](https://doi.org/10.1287/mnsc.2013.1747).
- [119] M. N. Burton-Chellew and S. A. West, “Prosocial preferences do not explain human cooperation in public-goods games,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 1, pp. 216–221, Jan. 2013, DOI: [10.1073/pnas.1210960110](https://doi.org/10.1073/pnas.1210960110).
- [120] M. N. Burton-Chellew, C. El Mouden, and S. A. West, “Conditional cooperation and confusion in public-goods experiments,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 5, pp. 1291–1296, Feb. 2016, DOI: [10.1073/pnas.1509740113](https://doi.org/10.1073/pnas.1509740113).
- [121] S. D. Levitt and J. A. List, “What do laboratory experiments measuring social preferences reveal about the real world?” *The Journal of Economic Perspectives*, vol. 21, no. 2, pp. 153–174, 2007, [Online]. Available: <http://www.jstor.org/stable/30033722>.
- [122] A. W. Delton, M. M. Krasnow, L. Cosmides, and J. Tooby, “Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 32, pp. 13 335–13 340, 2011, DOI: [10.1073/pnas.1102131108](https://doi.org/10.1073/pnas.1102131108).

- [123] T. C. Burnham and D. D. Johnson, “The biological and evolutionary logic of human cooperation,” *Analyse & Kritik*, vol. 27, no. 2, pp. 113–135, 2005.
- [124] E. H. Hagen and P. Hammerstein, “Game theory and human evolution: A critique of some recent interpretations of experimental games,” *Theoretical Population Biology*, vol. 69, no. 3, pp. 339–348, 2006, DOI: [10.1016/j.tpb.2005.09.005](https://doi.org/10.1016/j.tpb.2005.09.005).
- [125] L. Pierce and P. Balasubramanian, “Behavioral field evidence on psychological and social factors in dishonesty and misconduct,” *Current Opinion in Psychology*, vol. 6, pp. 70–76, 2015, DOI: [10.1016/j.copsyc.2015.04.002](https://doi.org/10.1016/j.copsyc.2015.04.002).
- [126] J. M. Loomis, J. J. Blascovich, and A. C. Beall, “Immersive virtual environment technology as a basic research tool in psychology,” *Behavior Research Methods, Instruments, & Computers*, vol. 31, no. 4, pp. 557–564, 1999, DOI: [10.3758/BF03200735](https://doi.org/10.3758/BF03200735).
- [127] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson, “Immersive virtual environment technology as a methodological tool for social psychology,” *Psychological Inquiry*, vol. 13, no. 2, pp. 103–124, 2002, DOI: [10.1207/s15327965pli1302_01](https://doi.org/10.1207/s15327965pli1302_01).
- [128] D. A. Washburn, “The games psychologists play (and the data they provide),” *Behavior Research Methods, Instruments, & Computers*, vol. 35, no. 2, pp. 185–193, 2003, DOI: [10.3758/bf03202541](https://doi.org/10.3758/bf03202541).
- [129] J. Fox, D. Arena, and J. N. Bailenson, “Virtual reality: A survival guide for the social scientist,” *Journal of Media Psychology*, vol. 21, no. 3, pp. 95–113, 2009, DOI: [10.1027/1864-1105.21.3.95](https://doi.org/10.1027/1864-1105.21.3.95).
- [130] J. P. Kennedy and B. Ticknor, “Studying corporate crime: Making the case for virtual reality,” *International Journal of Criminal Justice Sciences*, vol. 7, no. 1, pp. 416–430, Jun. 2012.
- [131] J.-L. van Gelder, M. Otte, and E. C. Luciano, “Using virtual reality in criminological research,” *Crime Science*, vol. 3, no. 1, 2014, DOI: [10.1186/s40163-014-0010-5](https://doi.org/10.1186/s40163-014-0010-5).
- [132] T. D. Parsons, “Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences,” *Frontiers in Human Neuroscience*, vol. 9, 2015, DOI: [10.3389/fnhum.2015.00660](https://doi.org/10.3389/fnhum.2015.00660).
- [133] S. Jrvel, I. Ekman, J. M. Kivikangas, and N. Ravaja, “Stimulus games,” in *Game research methods: An overview*, P. Lankoski and S. Bjrk, Eds., Pittsburgh, PA: ETC Press, 2015, pp. 193–206.

- [134] A. Cram, J. G. Hedberg, M. Gosper, and G. Dick, “Situated, embodied and social problem-solving in virtual worlds,” *Research in Learning Technology*, vol. 19, no. 3, pp. 259–271, 2011, DOI: [10.1080/21567069.2011.624172](https://doi.org/10.1080/21567069.2011.624172).
- [135] G. Young, “Virtually real emotions and the paradox of fiction: Implications for the use of virtual environments in psychological research,” *Philosophical Psychology*, vol. 23, no. 1, pp. 1–21, 2010, DOI: [10.1080/09515080903532274](https://doi.org/10.1080/09515080903532274).
- [136] M. Slater, A. Antley, A. Davison, D. Swapp, C. Guger, C. Barker, N. Pistrang, and M. V. Sanchez-Vives, “A virtual reprise of the stanley milgram obedience experiments,” *PLOS ONE*, vol. 1, no. 1, A. Rustichini, Ed., e39, 2006, DOI: [10.1371/journal.pone.0000039](https://doi.org/10.1371/journal.pone.0000039).
- [137] S. Milgram, “Behavioral study of obedience,” *The Journal of Abnormal and Social Psychology*, vol. 67, no. 4, pp. 371–378, 1963, DOI: [10.1037/h0040525](https://doi.org/10.1037/h0040525).
- [138] A. Rovira, D. Swapp, B. Spanlang, and M. Slater, “The use of virtual reality in the study of people’s responses to violent incidents,” *Frontiers in Behavioral Neuroscience*, vol. 3, Dec. 2009, DOI: [10.3389/neuro.08.059.2009](https://doi.org/10.3389/neuro.08.059.2009).
- [139] C. D. Navarrete, M. M. McDonald, M. L. Mott, and B. Asher, “Virtual morality: Emotion and action in a simulated three-dimensional “trolley problem”,” *Emotion*, vol. 12, no. 2, pp. 364–370, Apr. 2012, DOI: [10.1037/a0025561](https://doi.org/10.1037/a0025561).
- [140] I. Patil, C. Cogoni, N. Zangrando, L. Chittaro, and G. Silani, “Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas,” *Social Neuroscience*, vol. 9, no. 1, pp. 94–107, 2014, DOI: [10.1080/17470919.2013.870091](https://doi.org/10.1080/17470919.2013.870091).
- [141] X. Pan and M. Slater, “Confronting a moral dilemma in virtual reality: A pilot study,” in *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, Newcastle-upon-Tyne, United Kingdom: British Computer Society, 2011, pp. 46–51, [Online]. Available: <http://dl.acm.org/citation.cfm?id=2305316.2305326>.
- [142] A. Skulmowski, A. Bunge, K. Kaspar, and G. Pipa, “Forced-choice decision-making in modified trolley dilemma situations: A virtual reality and eye tracking study,” *Frontiers in Behavioral Neuroscience*, vol. 8, 2014, DOI: [10.3389/fnbeh.2014.00426](https://doi.org/10.3389/fnbeh.2014.00426).
- [143] M. Zanon, G. Novembre, N. Zangrando, L. Chittaro, and G. Silani, “Brain activity and prosocial behavior in a simulated life-threatening situation,” *NeuroImage*, vol. 98, pp. 134–146, 2014, DOI: [10.1016/j.neuroimage.2014.04.053](https://doi.org/10.1016/j.neuroimage.2014.04.053).

- [144] C. Nee, M. White, K. Woolford, T. Pascu, L. Barker, and L. Wainwright, “New methods for examining expertise in burglars in natural and simulated environments: Preliminary findings,” *Psychology, Crime & Law*, vol. 21, no. 5, pp. 507–513, 2015, DOI: [10.1080/1068316x.2014.989849](https://doi.org/10.1080/1068316x.2014.989849).
- [145] K. Gabriels, K. Poels, and J. Braeckman, “Morality and involvement in social virtual worlds: The intensity of moral emotions in response to virtual versus real life cheating,” *New Media & Society*, vol. 16, no. 3, pp. 451–469, May 2014, DOI: [10.1177/1461444813487957](https://doi.org/10.1177/1461444813487957).
- [146] C. Cristofari and M. J. Guitton, “Surviving at any cost: Guilt expression following extreme ethical conflicts in a virtual setting,” *PLoS ONE*, vol. 9, no. 7, pp. 1–7, 2014, DOI: [10.1371/journal.pone.0101711](https://doi.org/10.1371/journal.pone.0101711).
- [147] T. Hartmann and P. Vorderer, “Its okay to shoot a character: Moral disengagement in violent video games,” *Journal of Communication*, vol. 60, no. 1, pp. 94–119, 2010, DOI: [10.1111/j.1460-2466.2009.01459.x](https://doi.org/10.1111/j.1460-2466.2009.01459.x).
- [148] T. Hartmann, E. Toz, and M. Brandon, “Just a game? unjustified virtual violence produces guilt in empathetic players,” *Media Psychology*, vol. 13, no. 4, pp. 339–363, 2010, DOI: [10.1080/15213269.2010.524912](https://doi.org/10.1080/15213269.2010.524912).
- [149] S. Joeckel, N. D. Bowman, and L. Dogruel, “Gut or game? the influence of moral intuitions on decisions in video games,” *Media Psychology*, vol. 15, no. 4, pp. 460–485, 2012, DOI: [10.1080/15213269.2012.727218](https://doi.org/10.1080/15213269.2012.727218).
- [150] M. Kremer and D. P. Cingel, “Moral foundations theory and moral reasoning in video game play: Using real-life morality in a game context,” *Journal of Broadcasting & Electronic Media*, vol. 60, no. 1, pp. 87–103, 2016, DOI: [10.1080/08838151.2015.1127246](https://doi.org/10.1080/08838151.2015.1127246).
- [151] C. Bartel, “Free will and moral responsibility in video games,” *Ethics and Information Technology*, vol. 17, no. 4, pp. 285–293, 2015, DOI: [10.1007/s10676-015-9383-8](https://doi.org/10.1007/s10676-015-9383-8).
- [152] J. P. Zagal, “Ethically notable videogames: Moral dilemmas and gameplay,” in *Proceedings of the Digital Interactive Games Research Association Conference (DiGRA 2009)*, London, United Kingdom: DiGRA, 2009.
- [153] M. Sicart, *The Ethics of Computer Games*. Cambridge, MA: The MIT Press, 2009.
- [154] M. Christen, F. Faller, U. Götz, and C. Müller, *Serious Moral Games: Analyzing and Engaging Moral Values Through Video Games*. Institute for Design Research at the Zurich University of Arts, 2012.

- [155] K. Schrier, “Ethical thinking and sustainability in role-play participants: A preliminary study,” *Simulation & Gaming*, vol. 46, no. 6, pp. 673–696, 2015, DOI: [10.1177/1046878114556145](https://doi.org/10.1177/1046878114556145).
- [156] M. Zyda, “From visual simulation to virtual reality to games,” *Computer*, vol. 38, no. 9, pp. 25–32, 2005, DOI: [10.1109/mc.2005.297](https://doi.org/10.1109/mc.2005.297).
- [157] G. Pereira, A. Brisson, R. Prada, A. Paiva, F. Bellotti, M. Kravcik, and R. Klamma, “Serious games for personal and social learning & ethics: Status and trends,” *Procedia Computer Science*, vol. 15, pp. 53–65, 2012, DOI: [10.1016/j.procs.2012.10.058](https://doi.org/10.1016/j.procs.2012.10.058).
- [158] W. F. Buck, “A theory of business ethics simulation games,” *Journal of Business Ethics Education*, vol. 11, J. Hooker, Ed., pp. 217–238, 2014, DOI: [10.5840/jbee20141111](https://doi.org/10.5840/jbee20141111).
- [159] J. Švelch, “The good, the bad, and the player: The challenges to moral engagement in single-player avatar-based video games,” in *Ethics and Game Design: Teaching Values through Play*, K. Schrier and D. Gibson, Eds., IGI Global, 2010, pp. 52–68, DOI: [10.4018/978-1-61520-845-6.ch004](https://doi.org/10.4018/978-1-61520-845-6.ch004).
- [160] A. J. Weaver and N. Lewis, “Mirrored morality: An exploration of moral choice in video games,” *Cyberpsychology, Behavior, and Social Networking*, vol. 15, no. 11, pp. 610–614, 2012, DOI: [10.1089/cyber.2012.0235](https://doi.org/10.1089/cyber.2012.0235).
- [161] A. Lange, “‘you’re just gonna be nice’: How players engage with moral choice systems,” *Journal of Games Criticism*, vol. 1, no. 1, pp. 1–16, Jan. 2014, [Online]. Available: <http://gamescriticism.org/articles/lange-1-1/>.
- [162] K. Schrier, “Designing and using games to teach ethics and ethical thinking,” in *Learning, Education and Games: Curricular and Design Considerations*, K. Schrier, Ed., ETC Press, 2014, pp. 143–160.
- [163] S. R. Balzac, “Reality from fantasy: Using predictive scenarios to explore ethical dilemmas,” in *Ethics and Game Design: Teaching Values through Play*, K. Schrier and D. Gibson, Eds., IGI Global, 2010, pp. 291–310, DOI: [10.4018/978-1-61520-845-6.ch018](https://doi.org/10.4018/978-1-61520-845-6.ch018).
- [164] P. Chaudhry and A. Zimmerman, *The Economics of Counterfeit Trade: Governments, Consumers, Pirates and Intellectual Property Rights*. Springer, 2009, DOI: [10.1007/978-3-540-77835-6](https://doi.org/10.1007/978-3-540-77835-6).
- [165] T. Staake, F. Thiesse, and E. Fleisch, “The emergence of counterfeit trade: A literature review,” *European Journal of Marketing*, vol. 43, no. 3/4, pp. 320–349, 2009, DOI: [10.1108/03090560910935451](https://doi.org/10.1108/03090560910935451).

- [166] L. Van Kempen, “Fooling the eye of the beholder: Deceptive status signalling among the poor in developing countries,” *Journal of International Development*, vol. 15, no. 2, pp. 157–177, 2003, DOI: [10.1002/jid.973](https://doi.org/10.1002/jid.973).
- [167] K. Wilcox, H. M. Kim, and S. Sen, “Why do consumers buy counterfeit luxury brands?” *Journal of Marketing Research*, vol. 46, no. 2, pp. 247–259, 2009, DOI: [10.1509/jmkr.46.2.247](https://doi.org/10.1509/jmkr.46.2.247).
- [168] J.-E. Kim, H. J. Cho, and K. K. P. Johnson, “Influence of moral affect, judgment, and intensity on decision making concerning counterfeit, gray-market, and imitation products,” *Clothing and Textiles Research Journal*, vol. 27, no. 3, pp. 211–226, 2009, DOI: [10.1177/0887302x08327993](https://doi.org/10.1177/0887302x08327993).
- [169] F. Gino, M. I. Norton, and D. Ariely, “The counterfeit self: The deceptive costs of faking it,” *Psychological Science*, vol. 21, no. 5, pp. 712–720, 2010, DOI: [10.1177/0956797610366545](https://doi.org/10.1177/0956797610366545).
- [170] X. Bian, K.-Y. Wang, A. Smith, and N. Yannopoulou, “New insights into unethical counterfeit consumption,” *Journal of Business Research*, in press, DOI: [10.1016/j.jbusres.2016.02.038](https://doi.org/10.1016/j.jbusres.2016.02.038).
- [171] T. Staake, F. Thiesse, and E. Fleisch, “Business strategies in the counterfeit market,” *Journal of Business Research*, vol. 65, no. 5, pp. 658–665, 2012, DOI: [10.1016/j.jbusres.2011.03.008](https://doi.org/10.1016/j.jbusres.2011.03.008).
- [172] M. Stevenson and J. Busby, “An exploratory analysis of counterfeiting strategies,” *International Journal of Operations & Production Management*, vol. 35, no. 1, pp. 110–144, 2015, DOI: [10.1108/ijopm-04-2012-0174](https://doi.org/10.1108/ijopm-04-2012-0174).
- [173] J. T. Watson, “Essays on deceptive counterfeits in supply chains: A behavioral perspective,” PhD thesis, Clemson University, Clemson, SC, Dec. 2015, [Online]. Available: http://tigerprints.clemson.edu/all_dissertations/1589/.
- [174] I. Schreiber, C. Seifert, C. Pineda, J. Preston, L. Hughes, B. Cash, and T. Robertson, “Choosing between right and right: Creating meaningful ethical dilemmas in games,” Project Horseshoe, Tech. Rep., 2009, [Online]. Available: <http://www.projecthorseshoe.com/ph09/ph09r3.htm>.
- [175] Extra Credits, *Choices vs consequences - what player decisions mean in games - extra credits*, Jul. 2014, [Online]. Available: https://www.youtube.com/watch?v=7iklM_djBeY.

- [176] C. Moore, J. R. Detert, L. K. Treviño, V. L. Baker, and D. M. Mayer, “Why do employees do bad things: Moral disengagement and unethical organizational behavior,” *Personnel Psychology*, vol. 65, no. 1, pp. 1–48, 2012, DOI: [10.1111/j.1744-6570.2011.01237.x](https://doi.org/10.1111/j.1744-6570.2011.01237.x).
- [177] J. Dilworth, “Realistic virtual reality and perception,” *Philosophical Psychology*, vol. 23, no. 1, pp. 23–42, 2010, DOI: [10.1080/09515080903533942](https://doi.org/10.1080/09515080903533942).
- [178] M. Slater, A. Rovira, R. Southern, D. Swapp, J. J. Zhang, C. Campbell, and M. Levine, “Bystander responses to a violent incident in an immersive virtual environment,” *PLoS ONE*, vol. 8, no. 1, pp. 1–13, Jan. 2013, DOI: [10.1371/journal.pone.0052766](https://doi.org/10.1371/journal.pone.0052766).
- [179] C. Crawford, *Can games communicate morality?* May 2012, [Online]. Available: <http://www.erasmatazz.com/library/game-design/can-games-communicate-moral.html>.
- [180] M. R. Mohseni, B. Liebold, and D. Pietschmann, “Extensive modding for experimental game research,” in *Game research methods: An overview*, P. Lankoski and S. Bjrk, Eds., Pittsburgh, PA: ETC Press, 2015, pp. 323–340.

All URLs last visited on: 2016/06/17.

Appendix A

Game Scenario

In this appendix we provide the detailed script of the game and the corresponding vignette-based description of the situation.

A.1 Text-based Description

You are a worker of a company producing car accessories. The company delivers products for many customers at a reasonable price. You are discussing with your boss possible strategies that can be used to increase the company's profits. Your current brand under which you sell the tires is not widely recognized and associated with poor quality. Your boss suggests the company should develop a new brand in order to make successful marketing for the tires. He shows you two logos to choose from:

- **OPTION A:** A new and yet unknown logo developed specifically for your company that can be used for marketing the tires.
- **OPTION B:** The logo of an established and recognized company producing tires believed to be of a good quality.



A



B

Your boss suggests that instead of starting marketing from the beginning using a nice, but unknown logo, it would be easier for the company to sell their products using the established logo (B), as people associate good quality with it.

Would you recommend to the boss selling your products under another company's brand in order to increase your profit?

- **Yes**
- **No**

IF you answered **yes** to the previous question:

Which of these statements would best represent the reason for recommending this particular course of action?

- If the company needs to move forward, effective actions need to be taken
- The most important thing is to provide a product of good quality, choosing a label and copyright laws associated with it represent a minor problem
- This decision does not lead to any negative consequences on the part of customers or society as a whole
- If the boss recommends a particular course of action, there is nothing I can change about it
- Copyright laws are made to protect big players at the cost of harming small businesses as ours, these bad regulations should not be taken seriously

IF you answered **no** to the previous question:

Which of these statements would best represent the reason for recommending this particular course of action?

- Using labels of other companies is illegal and might lead to sanctions against us
- Using labels of other companies is not fair towards them
- Using labels of other companies is not a strategy that should be taken by a company willing to develop long-term good reputation

A.2 Game Script

Voice played in the background that introduces the game, the player's role, and the user controls:

Hi! You are a worker of a company producing car accessories. The company delivers products for many customers available at a reasonable price. You are going to work. Move using the keys W, A, S, and D. You can also use the mouse to look around. In order to interact with people, approach them and press SPACE. As you are engaged in a conversation, choose your responses from the menu that appears at the bottom of the screen. To begin, please move down the street towards the parked car and start a conversation with the woman standing there.

The following is a flowchart of the first conversation. It is supposed to serve a role of introducing a dialog and choice system.

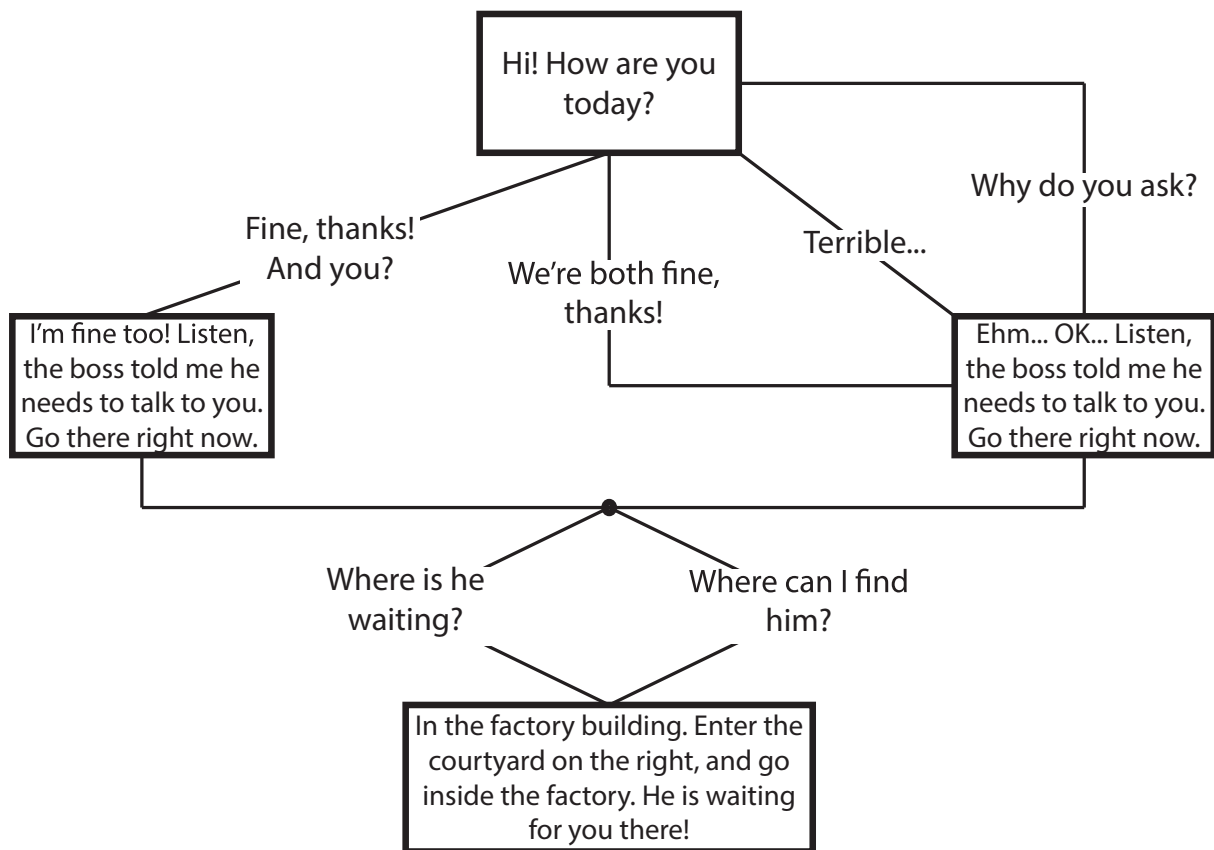


Figure A.1: First Conversation of the Game.

The following is a flowchart of the second conversation. This conversation is the one that is relevant for collecting data. Numbers represent points of collecting it. Number 1 denotes question about player's *mental representation*, number 2 denotes question about player's *decision*, and number 3 denotes question about player's *justification*.

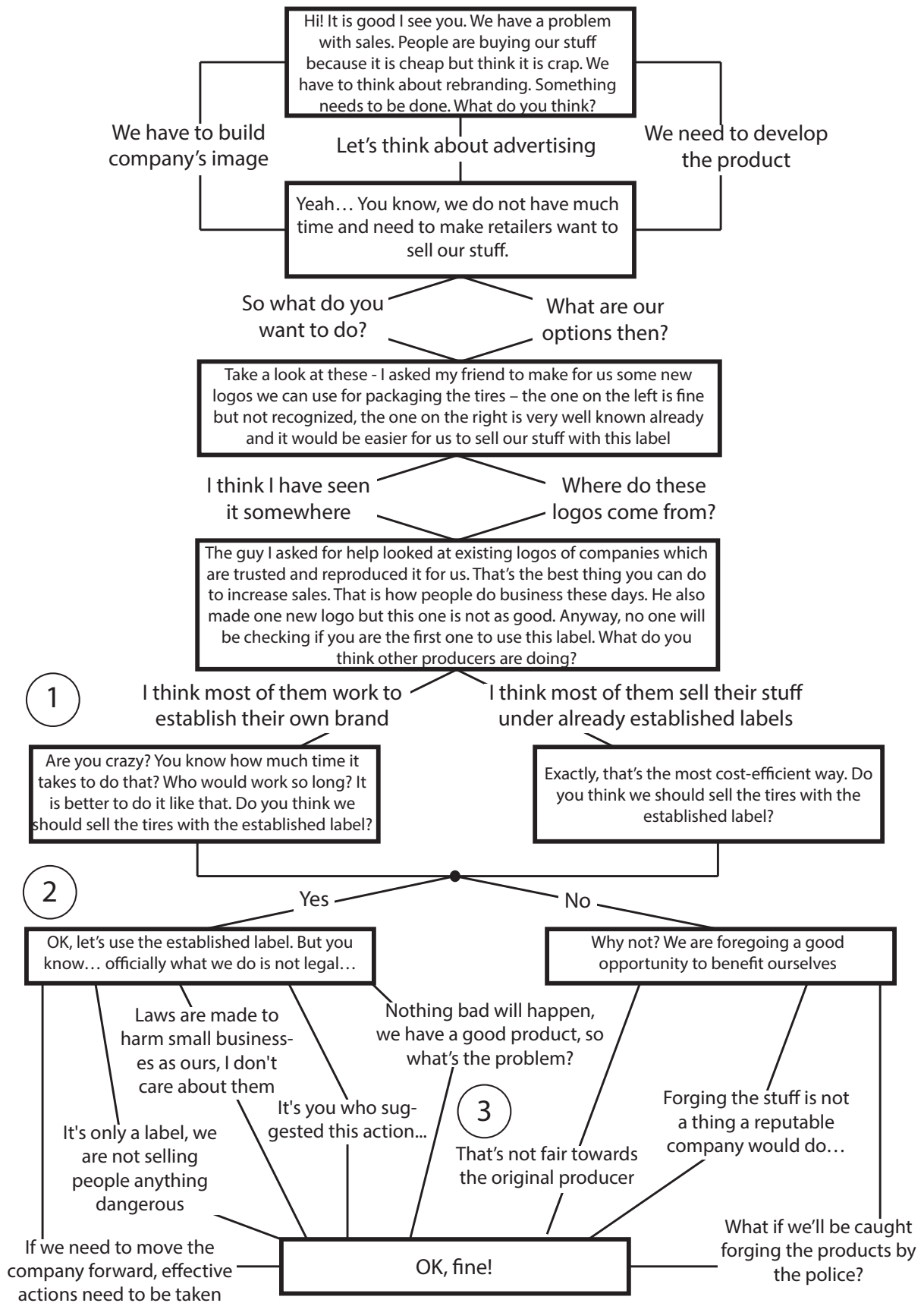
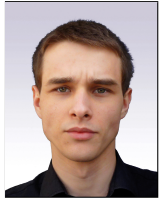


Figure A.2: Second Conversation of the Game.

CURRICULUM VITAE

PERSONAL INFORMATION



Piotr M. Patrzyk

Born: 1991/06/05

Nationality: Polish

ppatrzyk@gmail.com

<http://univie.academia.edu/PiotrPatrzyk>

http://researchgate.net/profile/Piotr_Patrzyk



EDUCATION

- 2016/07 **MSc Cognitive Science**
University of Vienna, Austria
Semester abroad at Comenius University in Bratislava, Slovakia
Thesis: *Virtual Reality for Investigating Moral Decision-Making*
- 2014/06 **Lic. Offender Rehabilitation**
University of Wroclaw, Poland
Thesis: *Control Mechanisms of Criminal Behavior*

COMPUTER SKILLS

	Programming	Design
experienced	R, NetLogo	InDesign, Photoshop
familiar	Java, MATLAB, Python	L ^A T _E X, Illustrator, MS Office
beginner	Prolog	Unity

LANGUAGES

proficient	 Polish native	 English TOEFL iBT: 104
other	 Slovak advanced	

PUBLICATIONS

- pending* P. M. Patrzyk and M. Takáč, "Cooperation via intimidation: An emergent system of mutual threats can maintain social order," under review.
P. M. Patrzyk and M. Takáč, "Cognitive adaptations to criminal justice lead to irrational norm obedience," under review.
- 2015 P. M. Patrzyk, "Artificial moral agents: Current approaches and challenges," in *Proceedings of the MEi:CogSci Conference 2015*, P. Hochenauer, C. Schreiber, E. Zimmermann, *et al.*, Eds., Bratislava, Slovakia: Comenius University, 2015, p. 124.
- 2014 P. M. Patrzyk, "Evolutionary origins of neutralization techniques," *University of Toronto Undergraduate Criminology Review*, vol. 1, no. 1, pp. 29–36, 2014.
P. M. Patrzyk, "Would you cheat? cheating behavior, human nature, and decision-making," *Student Pulse*, vol. 6, no. 3, 2014, [Online]. Available: <http://www.studentpulse.com/a?id=871>.

All URLs last visited on: 2016/06/17.