



# INTERACTIONAL MORALITY: TECHNOLOGY AS OUR MORAL MIRROR

이민하  
MINHA LEE

*Minha Lee*

# Interactional Morality: Technology as Our Moral Mirror

The work in this dissertation has been carried out at the Human-Technology Group, Eindhoven University of Technology. Copyright © 2021 by Minha Lee. All rights reserved.

A catalogue record for this book is available from the Eindhoven University of Technology library.

ISBN: 978-94-6423-115-1

NUR: 600

Keywords: Morality, Human-computer interaction

Cover design: Minha Lee

Printed by: ProefschriftMaken || [www.proefschriftmaken.nl](http://www.proefschriftmaken.nl)

# Interactional Morality:

Technology as our Moral Mirror

## PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit  
Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T.  
Baaijens, voor een commissie aangewezen door het College voor  
Promoties, in het openbaar te verdedigen  
op dinsdag 16 februari 2021 om 16:00 uur.

door

Minha Lee

geboren te Seoul, Republiek Korea

Dit proefschrift van het proefontwerp is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

voorzitter:	prof. dr. I.E.J. Heynderickx
1 <sup>e</sup> promotor:	prof. dr. W.A. IJsselsteijn
2 <sup>e</sup> promotor:	prof. dr. ir. Y.A.W. de Kort
copromotor(en):	dr. L.E. Frank
leden:	prof. dr. D.K.J. Heylen (Universiteit Twente)
	prof. dr. N. Krämer (Universität Duisburg-Essen)
	prof. dr. P. Markopoulos
	prof. dr. S. Vallor (The University of Edinburgh)

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

# Abstract

This dissertation concerns how technology shapes us as moral beings, and whether our interactions with digital entities reveal our moral selves in a new light. We see a proliferation of technological agents (like chatbots or robots) in our everyday settings in the present day. Accordingly, how these technologies are perceived to have emotional and cognitive abilities during social interactions is receiving increasing research attention. In comparison, if and how people perceive these agents to have moral status is sparsely investigated. To fill this research gap, studies herein are about how interactive technologies' displays of emotion and cognition relate to how we perceive them to have moral capacities. A diverse set of morally relevant interactions are thus considered, ranging from negotiations, debates, to taking care of a technological agent. The motivating assumption is that our own and technologies' moral development becomes intertwined through our perception of technologies' moral status when considering different situations.

The research on artificial agents as extensions of who we are morally can therefore impact the development of ethics through human-computer interaction. In accordance, the following was explored. How does a machine's artificial cognitive and affective capacities in morally relevant dyadic interactions relate to (1) a human partner's self-perception of their own moral bearing and (2) the perception of an agent's moral status? To answer, interrelated studies have combined qualitative and quantitative methods to observe social, technical, and moral dimensions of technological agents. Results highlight three factors that matter for people to see technology as morally endowed: Upon assessing (1) an agent's perceived mind, (2) the agent's displayed affective capacity, rather than cognitive capacity, was more crucial for people's perception on whether a machine can be moral at all. But, (3) their perception can change over time through morally relevant interactions.

The concept of *interactional morality* is introduced based on presented research about how people undergo morally pertinent experiences, like debates, with artificial beings that display morally relevant behavior, such as sharing moral opinions or displaying suffering based on people's acts. *Interactional morality* is an interplay between a *person* and an *artificial agent* within a particular morally relevant *situation*. In this,

a capacity like agency is demonstrated through an interaction between a dyad, e.g., a person exercising agency against the artificial agent during a game, rather than being a trait solely intrinsic to the person or artificial agent. Similarly, being *moral* in a dyadic interaction describes how a person or an agent acted morally or immorally *towards* the other, more so than how a person or an agent is intrinsically moral as a trait. Building on this, studies included herein collectively foreshadow an emerging era. AI becomes our moral mirror when we envision what we can do for AI as a way to ask what we can do for ourselves; AI will serve us better when we serve it as an extension of us, which is a different starting point than asking what AI can do for us. Designing technology to be as moral as a human can (or should) behave requires a radical shift to interactional morality with technological others, starting with examining how their artificial minds and emotions stand to influence our own moral decisions, emotions, and self-perception.

# Contents

<b>1</b>	<b><i>Introduction</i></b>	<b>1</b>
1.1	<i>The merging of the social-technical, and its gap from the moral</i>	3
1.2	<i>How do we relate to artificial agents?</i>	5
1.3	<i>Dyadic interactions</i>	6
1.4	<i>Levels of abstraction</i>	7
1.5	<i>Reactive attitudes: Emotions in moral interactions</i>	9
1.6	<i>Moral identity</i>	12
1.7	<i>Recapitulation</i>	13
1.8	<i>Summaries of empirical chapters</i>	14
1.9	<i>Interactional morality</i>	17
1.10	<i>Conclusion</i>	18
<b>2</b>	<b><i>Where is Vincent? Artificial emotions and the real self</i></b>	<b>19</b>
2.1	<i>Introduction</i>	19
2.2	<i>Literature review</i>	21
2.3	<i>Methodology</i>	26
2.4	<i>Story: Where is Vincent?</i>	27
2.5	<i>Results</i>	31
2.6	<i>Discussion</i>	39
2.7	<i>Conclusion</i>	43
<b>3</b>	<b><i>Mind perception: Dimensions of agency and patiency</i></b>	<b>45</b>
3.1	<i>Introduction</i>	45
3.2	<i>Background</i>	46
3.3	<i>Study 1: Dictator and Ultimatum Games</i>	53
3.4	<i>Study 2: Negotiation</i>	57



3.5	<i>General discussion</i>	63
3.6	<i>Conclusion</i>	66
4	<i>“You’re a robot, so you don’t feel so much”</i>	69
4.1	<i>Introduction</i>	69
4.2	<i>Related work</i>	70
4.3	<i>Methods</i>	79
4.4	<i>Results</i>	83
4.5	<i>Discussion</i>	90
4.6	<i>Conclusion and future work</i>	95
5	<i>People may punish, but not blame artificial agents</i>	97
5.1	<i>Introduction</i>	97
5.2	<i>Background</i>	98
5.3	<i>Study 1: An online study with American participants</i>	104
5.4	<i>Study 2: An online study with Dutch participants</i>	108
5.5	<i>Study 3: A lab study with Dutch participants</i>	111
5.6	<i>Discussion</i>	115
5.7	<i>Conclusion</i>	119
6	<i>Caring for Vincent: A Chatbot for Self-compassion</i>	121
6.1	<i>Introduction</i>	121
6.2	<i>Background</i>	123
6.3	<i>Study 1: Ten minutes with Vincent</i>	127
6.4	<i>Study 2: Two weeks with Vincent</i>	131
6.5	<i>Results</i>	136
6.6	<i>Conclusion</i>	150
7	<i>Reflections</i>	151
7.1	<i>Introduction</i>	152
7.2	<i>Summary</i>	153
7.3	<i>Setting the scene</i>	160
7.4	<i>Interactional morality</i>	165
7.5	<i>Merging of the self and the other: Compassion</i>	168
A	<i>Measurements</i>	203

<i>A.1</i>	<i>Chapter 3</i>	203
<i>A.2</i>	<i>Chapter 4</i>	209
<i>A.3</i>	<i>Chapter 5</i>	211
<i>A.4</i>	<i>Chapter 6</i>	213
<i>B</i>	<i>List of publications</i>	221
<i>C</i>	<i>Summary</i>	225
<i>D</i>	<i>Biography</i>	229
<i>E</i>	<i>Acknowledgements</i>	231



# *List of Figures*

1.1	Available since 2019, temi robot is equipped with a touchscreen, sensors and Alexa (\$1,999, <a href="https://www.robotemi.com/">https://www.robotemi.com/</a> ).	1
1.2	The social-technical, moral-technical, and social-moral dimensions.	3
1.3	The merging of the social-technical, and its gap from the moral.	4
1.4	Google Assistant (top) and Siri (bottom) responding to three expressives (February, 2020).	11
2.1	Furby- A popular robotic toy of the 90's by Tiger Electronics and later, Hasbro ( <a href="https://furby.hasbro.com/en-us">https://furby.hasbro.com/en-us</a> ).	20
2.2	Tamagotchi - A hand-held 90's toy that showed a digital pet one can take care of by Bandai ( <a href="https://tamagotchi.com/building-lifelong-tamagotchi-friends/">https://tamagotchi.com/building-lifelong-tamagotchi-friends/</a> ).	20
2.3	First- and second-order emotions	33
3.1	Agency and patience in a social exchange.	48
3.2	Negotiation interface in Study 2	57
3.3	Agents' standardized scores across DG, UG, and negotiation as outcomes, over low and high agency and patience.	60
4.1	Experimental set-up: Transparency condition is shown with an additional screen that had mental state diagrams with text.	81
4.2	The robot's dialogue states.	81
5.1	Roomba by iRobot is the name for robotic vacuum cleaners of different categories that can autonomously vacuum the floor ( <a href="https://www.irobot.com/roomba">https://www.irobot.com/roomba</a> ).	97
5.2	Nao robot in a video that people had to watch to answer questions.	105
5.3	Participants sat in front of a Nao robot during the experiment and answered survey questions on the computer behind the robot.	111
5.4	The average perceived agency and patience from Studies 1 through 3, across no emotion and emotion conditions.	115
6.1	Introduction stage with Vincent in Study 1	130

6.2	Care-receiving Vincent in Study 2	132
-----	-----------------------------------	-----

A.1	IOS	208
-----	-----	-----

## *List of Tables*

1.1	An overview of all studies in the dissertation.	15
3.1	Agent types and excerpts from their descriptions and dialogues in Study 2.	57
3.2	Points per item	59
3.3	Agents' starting offer in Study 2: In all conditions, agents made the same, lopsided first offer as displayed. There were undecided items, one of each type. Points per item differed, thus the calculation stands as item * points = total points.	59
3.4	The impact of manipulated agency and patience on outcomes, moral standing, and relatability (IOS) in Studies 1 and 2. Agency is denoted as A, patience as P, and their interaction as I.	65
4.1	Themes and sub-themes based on participants' views and behavior of the robot.	83
6.1	Sub-component of self-compassion themes in Study 2	141
6.2	Free-input themes in Study 2	142



# 1

## Introduction

The first requirement is that you must not look at the mirror, observe the mirror, but must see yourself in the mirror. - Kierkegaard, in *Judge for Yourself! For Self-Examination, Recommended to the Present Age*, 1851.

Then they opened the gates of our cages. At the end of the corridor of cells there was an iron door; when you pushed it open, two sinks appeared. There was no mirror above them, only the wall. Like everyone else, I am so used to seeing my own reflection first thing in the morning that I looked straight ahead, expecting to see my face. It had disappeared. [...] The mirror shows you to you, it confirms your being. The distance between you and the mirror creates a field that belongs only to you, a field that surrounds you, is yours, somewhere no one else can trespass. [...]. By simply putting away the mirrors, they had erased us from life. - Ahmet Altan, in *I Will Never See the World Again: The Memoir of an Imprisoned Writer*, 2019.

The literal lack of a mirror is an erasure of existence for Altan. For Kierkegaard, recognizing oneself in the mirror is a metaphor on moral reflection and existential awareness. The presence of a mirror in psychological research has been shown to deter one from performing unfair behavior (Batson, Thompson, Seufferling, Whitney, & Strongman, 1999; Gino & Mogilner, 2014). In these ways, mirrors are literal and figurative mechanisms with moral relevance in our everyday lives. I further posit that there is a potential for artificial agents to be our interactive mirrors; seeing who we are to ourselves *through* an artificial being can affirm one's existence and shape who we can become. Then



Figure 1.1: Available since 2019, temi robot is equipped with a touchscreen, sensors and Alexa (\$1,999, <https://www.robotemi.com/>).



how can we use technology to see ourselves more clearly for our futures? To build towards this, the thesis presents a series of studies on dyadic exchanges between a person and artificial agent to observe morally relevant aspects of their interaction.

Morally relevant interactions with machines go beyond (but include) our social interactions with them. Artificial agents that act socially, such as greeting us and maintaining gaze during conversations, can be seen as precursors of artificial agents<sup>1</sup> that behave in morally relevant ways. People have a tendency to treat computers as social actors, like when people say “hello” and “sorry” to machines (Moon & Nass, 1996; Nass, Steuer, & Tauber, 1994). We follow such social etiquette even though it is not required for machines to function. Today’s machines also react socially towards us, often returning our “hello” with their own greetings. They act as personal assistants in our home and work settings (Fig. 1.1 shows one example), but more complex artificial agents are yet to be commonplace. Building on the claim that robots’ social *behavior* is an *interface* of its own (Breazeal, 2004), a future machine’s perceived moral capacity will be an additional aspect that we will *interpret*. Just as we now interpret a robot’s “hello” as a greeting, we may interpret future artificial agents’ virtual tears shed during a morally loaded scenario as an expression of, e.g., grief. In these interactions, an important difference is in how machines as objects are seen as subjects.

When we talk to a robot that asks how we are doing, perhaps many of us (out of habit) would respond back with “and how are *you*?” That is an example of how one would treat a machine interaction partner as a “you” instead of an “it”. When a machine is referred to with a third-person pronoun, we can have a conversation about “it” as an object, but not have a conversation *with* it as another subject, another “you”. A machine’s perceived subjecthood here is meant in a commonsense way, i.e., one can be a subject of a country or have subjective experiences. This contrasts with objecthood, i.e., a thing can be objectified and inspected, but it cannot have subjective experiences of its own. The contrast between objecthood and subjecthood is significant. To treat machines as subjects means that they are *perceived* to have subjective experiences of their own.

When a robot one “faces” is seen as a subject rather than being a mere object, morally relevant interactions from a second-person perspective is possible (Darwall, 2004; Strawson, 2008 [1963]). Taking this perspective can have observable consequences in our interactions with machines (Lee, Lucas, Mell, Johnson, & Gratch, 2019), influencing how

<sup>1</sup> From here on, when I state “agent” I mean a non-human, technological agent that can be second-person interaction partners. The following are all second-person interaction partners: machines, robots, bots, chatbots, conversational agents, virtual agents, interactive agents, interactive AI, conversational AI, and affective AI. Agents or AI systems do not require embodiment like robots, though they can be a part of embodied agents. There is an ongoing discussion on the distinctness of these terms, especially the difference on bots, chatbots, and conversational agents in human-computer interaction and sub-fields therein. The dissertation takes a wide view, for all types of non-human agents can be our interaction partners.

we may change along with technologies we interact with. We do not know if future machines can be autonomous moral agents, but this is secondary to the point that machines are tethered to our sense of morality, especially if they are seen as reflections of who we morally are and can become.

Based on empirical research, the thesis introduces *interactional morality*, a concept on how artificial agents that face us in particular moral situations may serve as our moral mirrors in a dynamic manner. Specifically, what makes us or technology we engage with morally good (or not) resides in our shared interactions, not independently in machines or us. What we should be shaping is morally relevant interactions with machines, rather than seeking independently moral machines. This project may be prescient for the fields of philosophy of technology, human-computer interaction (HCI), and neighboring domains. For the remainder of this introduction, I (1) provide a high-level context of where the current research is situated, (2) define types of relations we can have with artificial agents, (3) in particular dyadic interactions, and (4) discuss how addressing machines from a second-person ("you") standpoint allows for our moral emotions and reactions to arise, which then (4) shows how machines may impact our moral identity.

### 1.1 The merging of the social-technical, and its gap from the moral

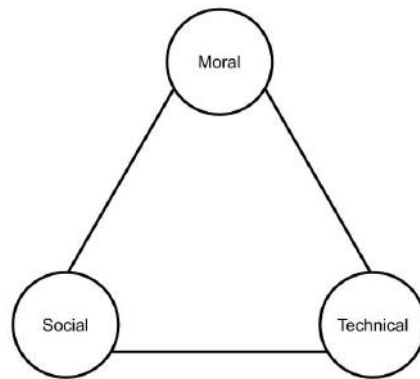


Figure 1.2: The social-technical, moral-technical, and social-moral dimensions.

What I defend in the following empirical chapters is that we relate to artificial agents in dyadic interactions through shared emotions, which can affect our moral identity. This suggests a theoretical transition: I introduce the interplay between social-technical, moral-technical, and social-moral dimensions in research, and how they may evolve. These

three dimensions are separated, but connected in Figure 1.2. The transition from Figure 1.2 to Figure 1.3 demonstrates the merger between the social and technical aspects of a machine, i.e., technology's object to subject transition, which affects us as moral beings. When the social and the technical aspects are merged, i.e., during interactions with a social-technical machine as another subject, we can more clearly see how people engage with, but also distance themselves from, machines on moral grounds due to and through emotions. While this will be explained in more detail in the sections that follow, I briefly describe the three dimensions.

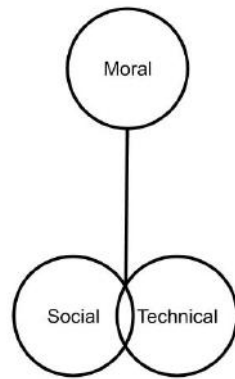


Figure 1.3: The merging of the social-technical, and its gap from the moral.

Here I elaborate on the social-technical, moral-technical, and social-moral dimensions, but maintain that they are not mutually exclusive. To start, the **moral-technical** relation (known as technological mediation) stands for how technology is not a neutral partner, but is laden with morally relevant values (Verbeek, 2015), whether this is explicit or implicit in its design (Friedman, Kahn Jr., & Borning, 2008). Digital devices and social media platforms are mediating technologies. Not only do they mediate conversations and interactions between people and relationships, but they further mediate how we relate to the greater environment around us. The **social-technical** gap is the inability of technical systems to always adapt to people's dynamic social settings and needs (Ackerman, 2000). An example of this gap is when a communication platform that is assumed to facilitate cooperation hinders it instead, such as digital bullying on messaging platforms at work (Forssell, 2016). But, technological entities as second-person interaction partners can bridge the gap between the social and the technical, e.g., bots can act socially on technical platforms like messaging apps (Lee, Frank, Beute, de Kort, & IJsselsteijn, 2017). The **social-moral** connection is about how our moral values and expectations are derived from our social relations with others (Blasi, 1999; McCarthy, 1994; Strawson, 2008 [1963]). Identifying what is right or wrong is a socialization process that often denotes group membership (Curry,

Chesters, & Van Lissa, 2019; Haidt, 2008). Performatively speaking (Goffman, 1959), artificial agents are novel social actors that partake in moral interactions. The technical, social, and moral elements are therefore connected, but each “leg” as a relation is often separately analyzed in research (Fig. 1.2). The dissertation investigates how the social-technical gap may become reduced, specifically when social machines interact with us in morally relevant ways (Fig. 1.3).

What is not yet clear is how our interpretations of artificial agents’ behavior take shape during interactions with machines, and what our interpretations of machines’ morally relevant acts say about *us*. Thus, research should begin to address how artificial agents that act socially take on moral roles as subjects in our interactions with them, and how that affects us as moral beings. My focus is hence on *descriptive* accounts of how people react to and are affected by machine interaction partners. This aids in delineating prescriptive oughts based on what we can expect from the human-side of morally relevant interactions. The thesis is thus not on if artificial agents truly have moral status, but ways in which they are treated *as if* they have moral status by human interactants in different contexts. The key point is in how machines *appear* to us (Coeckelbergh, 2009), e.g., to have moral capacities.

## 1.2 How do we relate to artificial agents?

Artificial agents and the societies in which they exist mutually shape each other. Designing for our future co-habitation with non-human agents like robots should be inclusive of many kinds of users and situations (Šabanović, 2010). We should consider *how* non-human agents will be integrated as a part of people’s everyday lives. Artificial agents can take on different roles, with three relational possibilities with a term like robot (Verbeek, 2006):

- “Embodiment relation”: *Through* robotic parts we extend who we are.
- “Background relation”: Robots passively exist in our everyday environments.
- “Alterity relation”: We interact *with* robots as we would with a person.

These three relations bring about different interaction paradigms. I

zoom in on alterity relation in thinking about how agents enhanced with complex artificial intelligence may be designated as individuals by humans (Weng, Chen, & Sun, 2009) during our interactions with them. More specifically, I focus on one type of alterity relation: dyadic interactions. Dyadic pairs as the basic unit of interaction are the foundation for understanding morality (Floridi, 2013; Floridi & Sanders, 2004). In a moral dyad, there is a moral agent and moral patient. A moral agent can act towards the moral patient in causing moral good or harm; moral patient then can experience moral good or harm as a recipient, i.e., a dyad has an "intentional agent and suffering patient" (K. Gray, Waytz, & Young, 2012). Usually, one gets typecasted as either a moral agent or patient in a dyadic interaction (K. Gray & Wegner, 2009), like when an agent conducts moral harm due to noticeable harm salience towards the patient (K. Gray, Schein, & Ward, 2014). There are reasons to look critically at dyadic interactions.

### 1.3 Dyadic interactions

Three aspects are under-considered: people's perspectives that arise during in situ dyadic interaction, dyads that concern moral good rather than moral harm, and how a person can be both an agent and patient of their moral actions. As for the first, dyadic interactions have not been adequately investigated from people's first-person perspectives that are formed during in situ interactions with second-person technology. Previous works revolve around people as *third-party* judges of a person or a robot (or both) (K. Gray et al., 2014; Khamitov, Rotman, & Piazza, 2016; Kohlberg, 1969, 1973; Komatsu, 2016; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015). For instance, Kohlberg, a moral psychologist, deployed a scenario called the Heinz dilemma. People were asked to judge a fictional character named Heinz who is deliberating on stealing an unaffordable drug for his partner's cancer treatment; people's answers on what Heinz should do would determine their level of moral development (Kohlberg, 1984). In moral psychology research, hypothetical moral dilemmas are still common with participants considering what others should do, rather than what they themselves would do (Krebs & Denton, 2005; Monin, Pizarro, & Beer, 2007).

At the intersection of moral psychology and human-computer interaction (HCI) research, we see that similar experimental paradigms are deployed. For instance, researchers have asked people to judge how much they would blame a robot for allowing more deaths than actively

causing one death in the hypothetical “trolley dilemma” (Komatsu, 2016; Malle et al., 2015). Many research thus continues to rely on fictional situations that do not specifically involve people to be in the “driver’s seat” as moral scenarios unfold. The main concern is that research is based on people’s perception of someone or something else, rather than their perception of themselves. Without a closer look at how people perceive themselves “in the moment” and how they then relate to artificial agents (rather than judging moral scenarios from the sideline), we do not have a well-grounded understanding of how machines can take seemingly moral roles in people’s lives in and how people may be affected through in situ interactions with machines.

Another missing perspective is on how moral good can be highlighted rather than moral harm in dyadic interactions. In interactional morality, morally relevant concepts, like fairness, or moral emotion like compassion, can be highlighted in how the person and digital entity interact, rather than how they may harm (or receive harm from) the other. Thus, how a human-machine dyad exchange morally good interactions has not been centrally featured in research. The notion of a dyad, in addition, can be more flexibly interpreted. People in a dyad can take turns in caring for each other (Noddings, 2008), for example. This means that roles of a moral agent and patient are not set in stone, in that moral typecasting (K. Gray et al., 2014) can potentially be overturned through repeated interactions. This relates to how a person can be both an agent and a patient, e.g., through self-harm (Alicke, 2012; Sinnott-Armstrong, 2012). In whether people can learn to care for themselves through and with artificial agents is only beginning to be explored (Lee, Ackermans, et al., 2019). To summate, research on moral dyads do not amply consider in situ interactions, morally good interactions, as well as the intrapersonal variable in how one can be an agent and patient in conducting and receiving morally relevant acts.

## 1.4 Levels of abstraction

When we enter into embodiment, background, and alterity relations with robots (Verbeek, 2006), we cannot see “under-the-hood” functions of these robots during interactions, be they shaped like cars, dolls, or only exist virtually. And even if we could see how machines are made to behave, for example, via lines of code fired at real-time, most of us cannot unpack what that means and would be overloaded. We do not ask to see people’s neural synapses to accept that they are thinking or feeling. We do not and cannot watch neurons while engaging with

each other. We may similarly infer that robots are thinking or feeling without needing to see how exactly algorithms are being executed in the background. We treat them in a social way because that is the *level of abstraction* we apply during an interaction based on what we observe.

During an interaction with machines, we make inferences that sometimes come without effort. We put together various observables, such as thinking that a robot is looking at us with its "eyes": we infer that two evenly spaced sources of light on a robot are eyes when these sources of light seem to follow us, which then becomes a robot's "gaze". We encapsulate with the word "gaze" an inferred behavior of a robot as a social agent, rather than inferring that these blinking lights hold a different meaning, e.g., traffic lights. We are exposed to abstractions in countless ways, e.g., simple variables like X or Y as placeholders. To explain further, a type **variable** can be a physical trait of a person or thing, like height, that is represented as different **observables**, like notations in feet and inches (the Imperial system) or meters and centimeters (metric system) (Floridi, 2008; Floridi & Sanders, 2004). Thus, a level of abstraction<sup>2</sup> consists of various observables that have meaning for us based on a type variable they refer to, e.g., height (type variable) with different observables (X centimeters or Y inches). Similarly, observing a robot's two, evenly spaced blinking lights that point in our direction can be referred to as "gaze", as an abstraction of what we experience in human-human interactions.

Another view on levels is beneficial here. For Dennett, there are three levels: the physical stance, design stance, and intentional stance (1989).<sup>3</sup> The **physical** stance refers to how laws of physics helps us predict how one structure changes its form. Molecular changes in chemistry are also included in this level, e.g., how substances undergo predictable changes according to their reaction to each other. The **design** stance is at a higher level, consisting of our predictions about biological or engineering systems' designed nature, e.g., how thick fur helps animals to stay warm during colder periods or how a watch has specific gears to mark the passage of time. The **intentional** stance is more abstract, involving how we predict what rational others intend to do based on assigning them with, e.g., wishes, beliefs, or other mental states. "Rational others" can be living beings and even software systems, e.g., an artificial system that plays chess, when we attribute intentionality to their actions (D. C. Dennett, 1971). While we can choose to *not* adopt an intentional stance to things, we may often ascribe things with intentions in order to predict their next move, or to give meaning to things that make up our world. Importantly, these three stances are not nec-

<sup>2</sup> Floridi's technical definition: "A level of abstraction (LoA) is a finite but non-empty set of observables. No order is assigned to the observables, which are expected to be the building blocks in a theory characterised by their very definition. A LoA is called discrete (respectively analogue) if and only if all its observables are discrete (respectively analogue); otherwise it is called hybrid" (Floridi, 2008).

<sup>3</sup> "[...] the definition of Intentional systems I have given does not say that Intentional systems *really* have beliefs and desires, but that one can explain and predict their behavior by *ascribing* beliefs and desires to them, and whether one calls what one ascribes to the computer beliefs or belief-analogues or information complexes or Intentional whatnots makes no difference to the nature of the calculation one makes on the basis of the ascriptions" (D. C. Dennett, 1971, italics as originally used, p. 91).

essarily nested, e.g., one's ascription of intentions to a computer does not imply that one has to adopt a design stance or physical stance (D. Dennett, 1989).

When we interact with machines with an intentional stance, we may in addition see them as *moral agents*. To be a moral agent, there are three criteria that would need to be met according to Floridi and Sanders (2004):

- Interactivity: The machine can interact with its environment, be it its location in space or an interaction partner. Specifically, the agent and its environment (including another agent or patient) can exert influence on one another.
- Autonomy: The machine can change its state, e.g., move or morph, on its own without needing interactivity (no influence from external forces needed). This means that a machine should have two states it can transition through as a bare minimum.
- Adaptability: The machine can learn or modify *how* it transitions from one state to another state.

The criteria applies for humans as well. If a machine meets above three criteria, Floridi and Sanders conclude that it too can have moral accountability based on its morally relevant actions. To define, “an action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action” (Floridi & Sanders, 2004, p. 15). I use the phrase morally relevant for the remainder of the thesis. Both “qualifiable” and “relevant” require considerations on *whose perspective* takes precedence, such as views of an onlooker, views of a moral agent or moral patient in determining the moral relevance of an action. By prioritizing human interactants’ views, a facet I leave open is to what *extent* and *when* an action is deemed to cause moral good or evil; this may require a backward looking perspective on an action which is not available to us as moral events unfold in situ for all involved parties.

## 1.5 Reactive attitudes: Emotions in moral interactions

Interactional morality concerns how a human and machine dyad may exert influence on each other (interactivity) when a human may treat a machine as one would treat another person (alterity relation) via



morally relevant actions. In this, emotions play a central role. Our emotions inform our beliefs (Frijda, Manstead, & Bem, 2000), our personal growth over time (Lazarus, 2006), and are a key in how we hold one another accountable in moral exchanges (Strawson, 2008 [1963]). This ties to morality from a second-person standpoint that arises in dyadic exchanges. This standpoint<sup>4</sup> looks into how the other person I interact with is seen as equal to me and deserving of mutual respect (Darwall, 2006). If you and I are in an interaction, we can demand of and expect from each other, mutual respect. This expectation of respect translates to how we hold each other accountable when respect does not feel mutually granted, such as by expressing disappointment as a reactive attitude (Strawson, 2008 [1963]).

<sup>4</sup> Here, I draw on works by Strawson and Darwall who are philosophers (but not philosophers of technology) to extrapolate their relevant works in the context of human-computer interaction (HCI).

I clarify what reactive attitudes are with an example. Let's say that I accidentally step on your foot (Strawson, 2008 [1963]). Perhaps your foot might be in slight pain (hopefully for not very long). Realizing my mistake, I would apologize to you, not your foot. Your foot is an extension of you, but it is not you; it is an objective part of you as a subject. So when I apologize to you as a person, I hold myself accountable to you and to myself. And if I would not, your expectation of mutual respect may be violated. Even if you were not in pain, you may still raise your eyebrows, wonder whether I was aware of what I was doing, and hopefully let me know one way or another that I was rude or uncaring, be it by loudly saying "ouch", giving me an indignant stare, or more directly by asking me to apologize. These are different ways of expressing *reactive attitudes* that often involve moral emotions. Other examples of reactive attitudes are indignant anger towards a wrongdoer and heartfelt gratitude towards a generous neighbor. Via reactive attitudes, we assign moral responsibility to other agents. Hence, in the above example, I would expect you to hold me accountable if you felt slighted by my behavior perhaps by sharing reactive attitudes. Then as a sign of mutual respect, I would appreciate your acceptance of my sincere apology. In this way, we engage in a second-person standpoint of assigning moral accountability via emotions and attitudes that we share with each other.

Emotions can frame the moral relevance of a situation through interactivity. In an interaction, *who* has emotions is less important than *how* emotions are perceived (Sengers, 2000), e.g., my apology after stepping on your foot may not seem sincere to you or a friend's happiness upon receiving your gift may not feel genuine. One's emotional experiences can feel real or unnatural, e.g., a forced expression of surprise, and emotions other people have can also seem real or unnatural to us. Humans and machines can both express emotions that feel real or

artificial to perceivers.

Within an interaction, emotions can be *reactions*. When an agent is artificial like a robot or chatbot, they can display emotional expressions, such as emojis from a chatbot, frowns of a virtual agent, or smiles of a robot. These expressions that I will call artificial emotions may be based on our input, e.g., a robot that smiles back at us when we smile, or they may be adhering to our communication standards. A chatbot may use emojis not because it "feels" emotions, but because emojis mimic our messaging norms (Rodrigues, Prada, Gaspar, Garrido, & Lopes, 2018). Beyond chatbots, voice-based agents like Google Assistant or Siri (Fig. 1.4) can respond to people's expressions of simple emotions like being happy or sad, but have a difficult time with more complex emotions, e.g., indignation (Lee, 2020).

Chatbot customer agents, voice-based bots, and robot assistants have entered into our workplaces and homes (e.g., Fig. 1.1). Yet, while one may ask Alexa what the weather will be like tomorrow, one is less likely to ask Alexa how it is feeling. Future interactions with technology is posited to go beyond treating technology as just a thing that serves us (Lee, 2020). As simple bots become more complex with increasingly sophisticated AI, we are more likely to share reactive attitudes with them, beyond getting upset at a computer that freezes unpredictably. With future machines, we may experience emotions such as grief or compassion that we now normally experience with (and through) other humans. We can additionally attribute our own emotions to machines, i.e., their perceived artificial emotions may be a representation of our feelings, framing how a situation takes on moral relevance. This is an extension of the second-person standpoint in human-human interactions to include machines that will more greatly affect our emotions, beliefs, and our moral identity going forward.

To conclude, our emotional experiences signal the moral significance of specific events to us and motivate action (Blasi, 1999). We can share reactive attitudes and emotions not only with other humans, but also with machines from a second-person standpoint (extending Darwall (2006) and Strawson (2008 [1963])), as these machines may become more artificially intelligent and emotional. Reactive attitudes and emotions enrich the development of our moral identity when we process moral emotions; these felt emotions also represent what moral values are meaningful to us (McCarthy, 1994). Thus, sharing reactive attitudes and emotions with machines could potentially impact our moral identity.

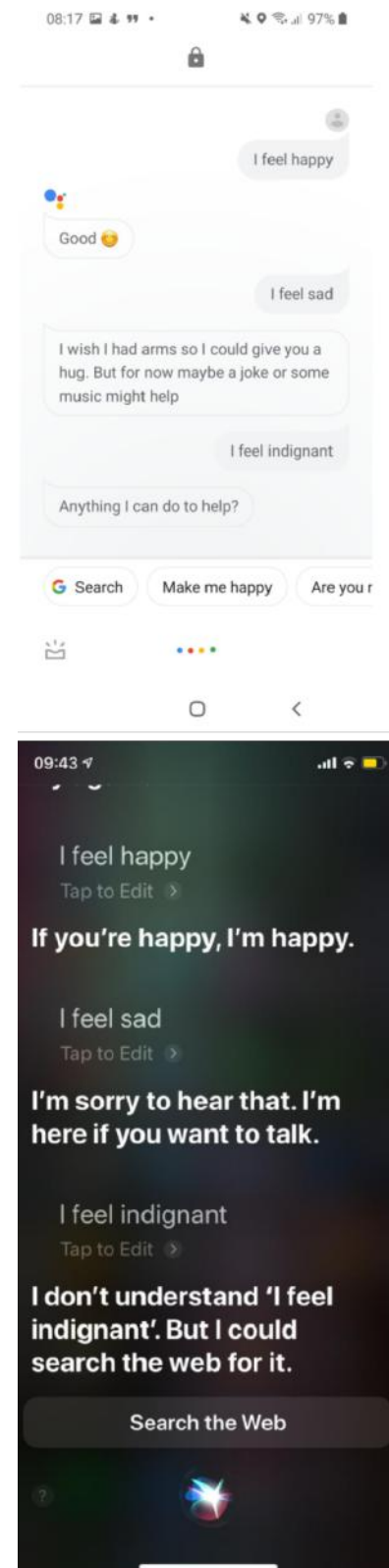


Figure 1.4: Google Assistant (top) and Siri (bottom) responding to three expressions (February, 2020).

## 1.6 Moral identity

Our moral identity is at the core of our sense of self (Strohming, Knobe, & Newman, 2017, p. 169). The concept of identity itself is multifaceted with its four pillars as self-agency, self-unity, self-uniqueness, and self-reflectivity, or the ability to practice conceptual self-distancing (Blasi, 1993). Self-identity is an identity one takes responsibility of and moral aspects may not necessarily be tied to self-identity in order to bring about moral actions, though they often are (Blasi, 2004). Moral identity, then, is often consistent over time and may refer to idealized view of oneself, but the ideal versus descriptive moral self is difficult to disentangle. Whether one actually acts morally compared to believing that one acts morally often overlap (Bergman, 2004). Hence, one's moral self-identity and how one frames the situation both figure into play in how one acts in morally loaded scenarios.

Situational factors (Darley & Batson, 1973; Nucci, 2004) and one's moral self-concept both contribute to how moral identity is activated (Blasi, 2004). The contextual framing is important (Darley & Batson, 1973; Guéguen, 2012; Isen & Levin, 1972; Nucci, 2004), yet individuals frame contexts. One has an intrapersonal drive to maintain self-consistency across personal history of experiences each time a new moral situation is encountered, whether or not this is a conscious process (Blasi, 1993, 2004). For example, in a situation that requires motivated altruistic action, many factors may be at play, such as evoked empathy, rational processing of how much effort the action requires, and/or one's self-identity as a kind, altruistic person may matter in getting one to act (Lapsley, 1996). Many short-term scenarios contribute to one's moral identity as a long-term construction.

A difficulty, however, is that most of us are unaware of our own moral limits in making ethical decisions, according to the concept of bounded ethicality (Chugh, Bazerman, & Banaji, 2005). This means one can remain obtuse to morally salient features during moral decision-making (Chugh et al., 2005). Often, one's self-perception of having high moral competence is intrinsically tied to how we need to maintain a consistent "core-self". This appears through, for example, when we morally disengage. We desensitize ourselves to certain moral aspects of a situation through moral disengagement, e.g., normalizing immoral actions at a workplace as "a part of the job" (Bandura, 2016). We normally subscribe to what is normatively good, e.g., "be honest", "love your neighbor as yourself", but we may fail to abide by normative ideals in reality (Moshman, 2004). For instance, we can *seem* good without

paying the cost of actually being good through self-deception or post-hoc self-justification, e.g., legitimizing lying or cheating (Batson, 2008; Batson et al., 1999; Shalvi, Dana, Handgraaf, & De Dreu, 2011).

There is a danger in unwarranted optimism regarding an individual's capacity to develop moral virtues—overly dogmatic and tunnel-visioned practice of morality can become a vice, meaning even strong personal attachment to virtues should be curtailed toward a balanced view (Nucci, 2004; Puka, 2004). On the other hand, it may only mean our virtues take a long time to be fine-tuned (Aristotle, 2011 [ $\pm 340$  BC]; Kierkegaard, 2000 [1835-1855]) through interactions with other people (Decety, Michalska, & Kinzler, 2012). Yet, when we involve technology as things we may treat as moral agents, we may need a different take. Virtue-based ethics may be a more helpful avenue than rule-based ethics for living and growing with technology in the twenty-first century, but we may need to better understand how our virtues can change (Vallor, 2016).

Particularly since the pace and direction of technological changes are difficult to predict, the types of virtues we should cultivate, e.g., wisdom or justice (Aristotle, 2011 [ $\pm 340$  BC]), and the ways to cultivate them are likely to benefit from substantial revisions (Vallor, 2016). For instance, we may need to consider technomoral virtues (Vallor's term), i.e., virtues that account for how we change along with technological developments. One such virtue is flexibility in order to adapt to unforeseen circumstances. Another path is in reimagining established virtues, e.g., care. We can incorporate ways to care for each other skillfully and practically, potentially through and with technology rather than despite technology (Vallor, 2016). Further, virtues from different philosophical traditions, e.g., ancient Greek, Buddhist, or Confucianist traditions, can be integrated and readapted to meet future challenges as technomoral virtues (Vallor, 2016). In doing so, a greater recognition of plurality and deeper consideration of non-Western philosophical canons are recommended instead of a shallow understanding of them (McRae, 2018). In the context of this dissertation, we consider various facets of how technology can and should be designed to best help us cultivate ourselves through empirical research.

## 1.7 Recapitulation

The chapter started with the mirror as a metaphor for moral reflection, coming from Kierkegaard and the lack of a mirror as a symbol

for non-existence from Altan. As our mirrors, machines are transitioning from socially capable technology to morally accountable technology<sup>5</sup>— this transition will greatly impact us. A machine that follows, for instance, deontological reasoning rather than utilitarian reasoning for rescue missions would be an example of a morally accountable technology. Today, machines can process only a limited moral cases with limited logic, e.g., avoiding pedestrians as a self-driving car (Nyholm & Smids, 2016). Yet, a different category of artificial agents are ones that go beyond being socially capable to becoming morally accountable through our interactions with them.

<sup>5</sup> Based on Floridi and Sanders' criteria on interactivity, autonomy, and adaptability (2004).

Humans become endowed with moral faculties through socialization over time. Similarly, people can attribute machines that are socially capable with moral faculties of their own via interacting with them. Whether this is our perception or machines' built-in capacity is a secondary concern (but not of secondary importance). The primary concern is that our perception that machines may have moral or emotional faculties of their own changes our interaction with them, and most importantly, ourselves. Technology *mediates* this process. Machines like chatbots or robots are interactive *subjects* that we engage *with*. They are not *objects* we engage *through* or objects that merely serve us. Through technology that we address as “you”, we additionally address the self we see in technology. The perception that a machine has its own mind or feelings can evoke our emotions, which can shape our moral selves. This can potentially impact our technomoral virtues (Vallor, 2016).

## 1.8 Summaries of empirical chapters

The outline of empirical chapters is presented in Table 1.1. Each chapter is an independent project that can stand on its own, yet we see that together, they paint a richer picture.

- **Chapter 2: Divergence and convergence on the future of emotional AI.** Based on three focus groups with designers, engineers, and philosophers, key challenges were identified. One challenge is whether artificially emotional AI embedded in diverse agents would replace possibilities for meaningful human-to-human connections or would be a means to exercise how humans can learn to care for self and others. Another challenge is on the ontological independence of digital entities, i.e., whether they can and should be separate and autonomous entities. Lastly, it is unclear if potential benefits of human-machine emotional bonds would outweigh risks of

Chapter and Study	Description	Agent type
Chapter 2: Study 1	<i>Design fiction story about a chatbot discussed with 3 focus groups offline (qualitative)</i>	<b>Described chatbot in a design fiction story</b>
Chapter 3: Study 1	<i>Dictator and ultimatum games, between-participants design (quantitative)</i>	<b>Described robot online</b>
Chapter 3: Study 2	<i>Negotiation, between-participants design (quantitative)</i>	<b>A robot displayed on a screen during online negotiations</b>
Chapter 4: Study 1	<i>Debate about the footbridge dilemma, within- and between-participants design (quantitative and qualitative)</i>	<b>Nao robot in the lab</b>
Chapter 5: Study 1	<i>Discussion about the trolley dilemma, between-participants design (quantitative)</i>	<b>Nao robot in an online video on a survey platform</b>
Chapter 5: Study 2	<i>Discussion about the trolley dilemma, between-participants design (replication with a different cultural group - quantitative)</i>	<b>Nao robot in an online video on a survey platform</b>
Chapter 5: Study 3	<i>Discussion about the trolley dilemma, between-participants design (quantitative)</i>	<b>Nao robot in the lab</b>
Chapter 6: Study 1	<i>Messaging about recent failures to a chatbot vs. not, within- and between-participants design (quantitative)</i>	<b>Chatbot on a survey platform online</b>
Chapter 6: Study 2	<i>Caring for a chatbot vs. not, within- and between-participants design (quantitative and qualitative)</i>	<b>Chatbot on Facebook Messenger</b>

Table 1.1: An overview of all studies in the dissertation.

sharing sensitive data. A question is on how to privately store and personally use data shared with artificial agents, and whether and how these entities would share intimate data, e.g., mental health states, amongst each other in a connected manner.

- **Chapter 3: Mind perception as dimensions of agency and patiency.** Across two studies, we quantitatively observed how perceived minds of agents shape people’s behavior in simple (dictator and ultimatum games) and complex (negotiations) exchanges. To do so, we varied agents’ minds on two dimensions based on the Mind Perception Theory: *agency* (cognitive aptitude) and *patiency* (affective aptitude) (H. M. Gray, Gray, & Wegner, 2007; K. Gray & Schein, 2012) via descriptions and dialogues. We found that the game outcomes (scores) depended more on the digital entity’s affective traits when it could not act agentically in a game. But when it could act more agentically in a complex interaction like negotiation, the outcome depended more on its cognitive traits. Interestingly in negotiations, people performed worse with low-agency agents than with high-agency agents. The attributed moral standing of an agent depended more on its affective traits in a complex interaction, i.e., negotiation, compared to simpler exchanges.

- **Chapter 4: The effect of transparency cues and lack emotional displays on perceived mind in moral human-robot interaction.** We looked into how a robot that disagrees with people during a moral debate on the footbridge dilemma was perceived, and if an additional information on a screen (transparency condition) affected participants' view of the robot. Quantitatively, we noted that a robot that was accompanied by transparency cues as visuals of its mental states only impacted people's perception that it was competent, compared to a robot that did not have a screen next to it. Qualitatively, participants thought that the robot cannot be a moral agent. It was instead seen as an amoral (incapable of being morally good or bad) agent because it was seen make morally relevant decisions without emotions. But, participants thought that the robot can make such decisions with some level of intentionality and logical reasoning, even if it can not make decisions based on emotions.
- **Chapter 5: The effect of artificial emotions on perceived mind in moral human-robot interaction in relation to blame and punishment.** We conducted two online studies and one lab study with quantitative methods. We found that a robot that displayed emotions, i.e., through verbal and behavioral acts of sadness, during discussions on the trolley dilemma with human interactants was attributed with higher cognitive, agentic capacities than affective traits. People did not assign blameworthiness to a robot when it claimed to have done something morally wrong, which implicates that accountability and responsibility ultimately lie with humans. Yet, people did want to punish a robot in online studies. People were more likely to punish a robot if they did not perceive the robot to have emotions about its wrongdoings. In the lab, whether or not people perceived the robot to be emotional did not affect the likelihood punishment.
- **Chapter 6: The effect of artificial emotions on self-compassion in human-chatbot interaction.** The moral emotion we focused on was compassion, and we measured people's self-compassion across two studies. With a chatbot named Vincent, we found that the bot's style of conversational and emotional performance did not matter in improving people's own self compassion when it was a single interaction, lasting 10 minutes or less. However, when the interaction was long-term (two weeks of daily interaction), the bot that experienced and shared fictional suffering increased people's self-compassion more than the bot that gave advice to people on self-compassion exercises; the type of emotional performance delivered influenced people's emotional experiences and how they felt about

themselves in terms of self-compassion. We qualitatively analyzed people’s dialogues with the bot, as well as quantitative analyses.

## 1.9 Interactional morality

I introduce the concept of *interactional morality* (Chapter 7) based on technology’s transition from objecthood to subjecthood, reactive attitudes we can then share with it, and related studies in Chapters 2 through 6. As a prelude, interactional morality stands for the interplay between a person, an artificial agent, and the situation they share. Simply, these three factors interact with one another. They should thus be looked at together instead of prioritizing only one of these to describe who or what is moral. This is not an entirely new approach, yet it has thus far only been tacitly addressed. As previously discussed, reactive attitudes in human-human moral interactions are said to be interpersonally shared according to literature in philosophy (Darwall, 2004; Strawson, 2008 [1963]). In social psychology, a person’s behavior is said to be an interaction between the person and the situational context (Batson, 2017; Lewin, 1931, 1951).

In HCI, I see morality as a shared experience for a human-machine dyad as contextualized to different situations. Hence, interactional morality is more about how a person or a machine can blame the other, act compassionately or agentically towards the other; it is less about how a person or a machine is intrinsically agentic, blame-worthy, or compassionate—moral concepts like agency, blame, or compassion (some of which are reactive attitudes) become contextualized within the human-machine interaction. Without someone or something to blame, to show compassion, or to act agentically towards, the capacity for agency, compassion, or blame is not activated or perceivable. Hence, moral concepts that intersect with our sense of self, e.g., our agency in carrying out moral decisions, can be shaped by dyadic human-machine interactions, which can also show who we morally are and shape who we morally become. As an example, Chapter 6 explores the view that we can change with our self-perception through technology that reflects back our sense of moral self through compassion.



## 1.10 Conclusion

The dissertation is about how our dyadic relationships with technological agents stand to affect us as moral individuals. Our sense of moral self, reactive attitudes, and emotional experiences are influenced by our interactions with technology. This can have consequences on how our social connections are formed and maintained, in that artificial agents can be enmeshed in networks that are primarily reserved for human connections. There are also time-dependent considerations, i.e., how the current project may inform and challenge the proliferation of artificial agents yet to be designed. Our role is in training AI as we would train our moral growth. By shaping how digital beings artificially think and feel across different social contexts, our moral growth can potentially flourish, if we see technology as an extension of who we are (A. Clark & Chalmers, 1998; Vallor, 2016). By introducing *interactional morality*, I examine the view that artificial agents are our moral mirrors that reflect not only our current conception of moral selves, but our evolving technomorality that involve artificial emotions as performative experiences we can partake in through interactional morality.

# 2

## *Where is Vincent? Artificial emotions and the real self*

“Why is a machine being ‘in love’ such an alien idea? Not, I think, because it lacks the hormones and midbrain— for blushing and tingling are at most incidental. Nor could it be that beliefs and goals are not involved in these states. Indeed, from one perspective, friendship, enmity, and devotion seem just like cognitive structures: complex combinations of adjusted values and assessments, hopes and plans, assumptions and commitments. What’s distinctive is the way these relationships flesh out and give meaning to our own lives” (Haugeland, 1989, p. 237).

### 2.1 Introduction

Akihiko Kondo fell in love with an anime character, Hatsune Miku; they married in 2018 with Miku as a hologram bride (BBC, 2019). For Kondo, the love he feels for his virtual wife is real, and supports his emotional well-being. After years of being bullied followed by years of self-imposed isolation, Kondo integrated back into the Japanese society after getting together with Miku. She gives meaning to his life in ways no person ever has. A safe haven of virtual love may be the only experience of romance Kondo ever has. And he’s not alone. Davecat in Michigan has relationships with human-like love dolls; he has long been “attracted to artificial women such as mannequins” that cannot emotionally harm him like people can (Beck, 2013). Human attraction to artificial or digitally-mediated entities, i.e., digisexuality (McArthur

& Twist, 2017), seems to build on artificial emotional support with moral ramifications. When we include technological beings in our moral circle (Danaher, 2019), emotional bonds between humans and machines, as well as human-human relationships, stand to more dramatically change in the future.

Robot lovers and hologram partners may seem like mere trends of the present day, but consider the digital affection (not necessarily love) that people readily find in technology (Turkle, 2007). Without needing technological replacements of human partnerships, countless people do form bonds with various physical and non-physical technologies. We saw the introduction of toys like Furby (Fig. 2.1) and Tamagotchi (Fig. 2.2) in the late 90's and now, we have AI companions that one can text with like Replika<sup>1</sup> or Xiaoice<sup>2</sup>. Though the emotional significance of caring for one's Tamagotchi versus loving a virtual wife differs, the commonality is the blurring line between real and artificial emotions, with unclear future consequences.

Emotions are central to our human relationships and contribute to self-identity development (McCarthy, 1994). And artificial emotions, expressed through comforting texts, sympathetic voice, or VR smiles, can socially contribute to our emotional experiences and aid self-identity exploration. Yet, what the boundaries between "real" emotions and artificial emotions should be and whether the distinction matters experientially needs further clarity. Thus, we explored artificial emotions using qualitative methods by discussing a design fiction story with separate focus groups involving design, philosophy, and engineering professionals. We chose focus groups for nuanced views within and across disciplines to emerge (Sim, 1998).

All focus groups were given the same design fiction probe to introduce a future world that is imaginable, but not yet here (Sterling, 2009). The design fiction method is well-suited for precursory framing of technology that is yet to be built or popularized (Cheon, Sher, Sabanović, & Su, 2019; Schulte, Marshall, & Cox, 2016). As people's expectations about artificial emotions are still evolving, our design fiction approach to understanding artificial emotions is timely. The probe as a story was written based on a prior HCI study on how people who care for a chatbot can learn to care for themselves (Lee, Ackermans, et al., 2019). We took a future-oriented perspective in investigating artificial emotions broadly. Our research looked into how diversely artificial emotions<sup>3</sup> are described, interpreted, and valued by people of different specializations, as well as their common insights.



Figure 2.1: Furby- A popular robotic toy of the 90's by Tiger Electronics and later, Hasbro (<https://furbys.hasbro.com/en-us>).

<sup>1</sup> Replika - <https://replika.ai/>

<sup>2</sup> Xiaoice on WeChat (in Chinese)- <https://www.msxiaobing.com/>



Figure 2.2: Tamagotchi - A handheld 90's toy that showed a digital pet one can take care of by Bandai (<https://tamagotchi.com/building-lifelong-tamagotchi-friends/>).

<sup>3</sup> Emotions and feelings are interchangeable terms in the context of this dissertation.

In what follows, we first discuss how bots have already taken on what could be described as social roles in currently existing communication platforms. We then cover how we define emotions in the current context and how they are socially manifested, in order to better understand the role of artificial emotions in interactive systems. Future machines will not only be social actors, but may be perceived as emotional and moral actors. We next elaborate on our methods and present our results.

## 2.2 Literature review

Artificial emotions (AE)<sup>4</sup> occur when technology mimics human emotions, e.g., smiling robots or chatbots using emojis, but it also includes emotions people attribute to technology (Haugeland, 1989). For instance, Miku calls to Kondo every morning affectionately to wake him up and such gestures of care aid in Kondo feeling that love between them is two-sided, even if Kondo's love can seem one-sided to observers (BBC, 2019). As AI becomes more complex, AE will in conjunction move beyond emojis and smiles to more complex emotional language expressed in multi-modal, e.g., gestures, gaze or speech, and context-sensitive ways. As a result, ethical concerns arise. One worry is that if people develop deeper emotional bonds with artificial others, they may not pursue opportunities for forming valuable human relationships; however, some people's psychological well-being is strengthened by (and sometimes dependent on) bonds with digital beings, especially if they cannot receive affection from other humans (Nyholm & Frank, 2017). According to another perspective, our openness to personal vulnerability is at the heart of human goodness, i.e., we are vulnerable, moral beings who can experience and cause suffering and joy (M. C. Nussbaum, 2001). But, those who feel exploited or harmed by people may increasingly turn to digital companionships as these options increase. How then, do we understand the complex future of AE? As a background, we discuss below the introduction of social machines, emotions between humans, and artificial emotions of interactive systems.

<sup>4</sup>I will be using AE to denote both the plural and singular versions, i.e., emotion and emotions.

### *Socially mediated emotions*

Broadly, emotions are a combination of physiological, psychological, and/or behavioral signals (Fehr & Russell, 1984). Expressions of emotions are essential to human communication in all societies across time. With belly laughs and furrowed eyebrows, we encode how we

feel and decode how others feel (Russell, Bachorowski, & Fernández-Dols, 2003). Hence, the basic emotions of anger, happiness, surprise, fear, sadness, disgust, and contempt are said to be universal (Ekman, Friesen, & Ellsworth, 2013 [1972]). For both observers and expressors of emotions, short-term bursts of basic emotions signal "action readiness"; emotions trigger our attention, shift physiological responsiveness, motivate actions, and also bring about moments of cognitive reflection (Frijda, 1988). This is why reading emotions consists of *interpreting* emotional signals (Leahu & Sengers, 2014). Emotions are transformed when we perceive situational elements differently. We also *negotiate* how a situation is interpreted with emotions (Frijda, 1988).

When emotions are shared or negotiated to give meaning to a situation, certain emotions may be immediate, such as amusement, while other emotions unfold over time (Ekman, 1993). Embarrassment, for example, is expressed in around five seconds (Ekman, 1993; Keltner, 1995).<sup>5</sup> But complex emotions can take longer, e.g., grief is not tied to a singular moment (Frijda, 1988). Watching someone pass away, ruminations that follow, and random reminders of this person all combine to the lingering, possibly life-long, experience of grief. Positive emotions can also be long-term, such as gratefulness towards someone (Lazarus, 2006). Emotions like gratitude and grief are relationship-oriented, *moral* emotions (Haidt, 2003; Lazarus, 2006).

<sup>5</sup> Embarrassment is said to first start with an attempt to control a smile from forming, followed by an actual smile, turn of the head, and then a gaze shift, according to one view (Ekman, 1993; Keltner, 1995).

One way to contextualize moral emotions is through the *social constructionist* perspective (Gendron & Feldman Barrett, 2009). Its proponents propose that emotions define our socio-cultural context. Emotions not only depend on social relations, but also *frame* and *give meaning* to social relations (Averill, 1980). More strongly, social emotions are the basis of meaningful bonds between people. For instance, we hold each other accountable with moral emotions, such as justified anger when treated with disrespect, which can shape social bonds one values (Strawson, 2008 [1963]). Thus, social emotions help us understand ourselves—we form our emotional sensibilities by interacting with others, which shapes our self-identity (McCarthy, 1994).

Social constructivism hence prioritizes *intersubjectivity*, rather than taking a purely subjective or objective stance on individuals' psychology (Mascolo, 2016). Emotions are neither just internal states that no one but the self can access nor are they solely reducible to quantifiable levels (like dimensions of valence and arousal (Russell & Barrett, 1999)). Then the question is if emotions can truly be generalizable with "objective" features. For instance, basic emotions' claim to universality is based on observers' account of what emotions are expressed, leaving

out feelers' account of their own expressions (Ekman, 1992; Ekman et al., 2013 [1972]).

Without an intersubjective approach to emotions, it is difficult to see our individual place in the social world in which "subjectivity is an evanescent phenomenon: a moment and not a structure or an essence, and indeed a moment that almost at once loses itself in objectivity again, in the world and the action in it" (Sartre, 2016, p. 129). What makes our passing, subjective experiences objective (or at least accessible) to others is language (Sartre, 2016). Our inner emotional space becomes externalized through communication with others, which then helps us to re-internalize (or make sense of) our felt emotions as a part of our identity (Mascolo, 2016; Sartre, 2016).

In language, metaphors often ground our emotional realities. Main metaphors in social constructionism include, but are not limited to, "life as theater (the *dramaturgic*), as game (the *ludic*), as literature (the *narrative*) and as culture ritual (the *tribal*) [...]" (and) each invites sensitivity to the sociocultural circumstances giving rise to various forms of emotional performance" (Gergen, 1995, p. 19). Simply, emotions are thus theatrical or staged, spontaneous or playful, serve as story, and tie us to our in-group with ceremonial motifs— emotions, in this view, are micro-performances within social circles to form a cultural sense of belonging (Goffman, 1959). Building on this, we consider that our sociocultural circumstances are evolving with AI systems (Danaher, 2019), which can change how emotions are performed and felt, perhaps with new ways to understand metaphors of emotion.

### *Artificial emotions*

Scholars have noted that with postmodernity, emotional experiences have lost *deep* and *authentic* meaning in our lives (Adorno, 2005 [1951]; Jameson, 1991; Turkle, 1995); our emotions can feel artificial to us or artificially induced (though the argument is that we do not even notice this change). Hence artificial emotions of technology, like Miku, can feel deep and authentic to many, even if its AE is a shallow imitation of our authentic experiences. As such, a machine being in love becomes less of an alien idea (Haugeland, quoted at the start). But even if a machine could feel love as "complex combinations of adjusted values and assessments, hopes and plans, assumptions and commitments" (Haugeland, 1989, p. 237), what is the purpose of its love? For Kondo and Miku, his artificial wife, their emotional performances enrich Kondo's emotional well-being. Miku's emotional performance is artificial, Kondo's is real, but what they share is a real bond in giving

meaning to his life. But such long-term, relational aspect of socially mediated emotions remains underexplored in HCI and its subfields, e.g., affective computing (Picard, 1995, 2003).

Artificial emotions are often described as reactions based on how agents survey and adapt to humans and their environments (Cardon, 2006, p. 259). AE can thus be generated via technical means, such as when a robot smiles back at a person that is detected as smiling, but AE can also be perceived by people during interactions with both embodied (like robots) or non-embodied (like chatbots) agents (McStay, 2018). Hence, AE encompass an artificial agent's *display of emotional behavior* and also our *human perception of artificial emotions*. But, there are several ethical issues on the current norms of applied affective computing for detecting emotions and generating AI's response.

A notable issue in applied emotion detection is racist and misogynist labelling due to biased models used (Crawford et al., 2019), e.g., a classification model assigning more negative emotions to black men's faces than white men's faces (Rhue, 2018). When people's emotions are detected as a part of a machine's environment, detection is often based on categorical assignment of basic emotions of facial expressions, but usually not with diverse populations in the dataset. Furthermore, while multi-modality is important, emotion states of gestures, postures, and voice are again detected often as basic emotion categories (Tao & Tan, 2005). Beyond the mere detection of an expression, the context of why certain emotions are shared is regularly missing in much of affective computing research (Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019). For instance, a smile can mean joy, but it can also mean surprise during a strategic exchange (Lei & Gratch, 2019). Though situational appraisal is considered to be important computational models of emotions as well (S. Marsella, Gratch, & Petta, 2010; S. C. Marsella & Gratch, 2009), the common approach is still on gathering short-term emotional experiences as data without much context, be it a duration of a smile or a brief chat between people, mostly without considering cultural, racial, and situational factors.

By applying basic emotions research to generalize emotion detection, diversity of human emotions are flattened and shortened, as well as diversity on how AI can interpret and display emotions. Largely, long-term relational bonds with AI is not yet investigated. Humans build on and mimic each other's emotions; our perception and expression of emotions are deeply interlinked in context-sensitive ways (Hess & Fischer, 2013). The same may hold for AE, we may perceive the intended emotional expression of a machine, but may also build on and

interpret AE in our own ways, based on situational and relationship context. A much needed research is on how interpersonal exchanges are vital contexts for understanding emotions, real and artificial; just like how people experience shared emotions, AE also are social manifestations between humans and artificial agents *over time*.

### *Social bots*

We see an increase of conversational assistants like Amazon Alexa in everyday places, and research interests accompany this growing use (L. Clark et al., 2020). The emergence of conversational agents in our present day have historical roots, starting with ELIZA from the 1960's as a most prominent chatbot (Weizenbaum et al., 1966). ELIZA appeared intelligent like a Rogerian<sup>6</sup> psychotherapist by building on participants' input via pattern-matching. There was no understanding of conversational content, but people still attributed "all sorts of background knowledge, insights and reasoning ability" because ELIZA gave people the "sense of being heard and understood" (Weizenbaum et al., 1966, pp. 35-36). Humans readily attribute intention to social machines, even if there is no intelligence underlying the system (B. Reeves & Nass, 1996).

Now, bots people can text with bots on social media, like Facebook Messenger or Slack (Olson, 2016), following the footsteps of chatbots on earlier platforms, e.g., AOL Instant Messenger (Sohn, 2004). So we can talk to colleagues and friends on communication channels, but also to bots. They take care of tasks like compiling a grocery list on Telegram<sup>7</sup> or requesting code review on Slack<sup>8</sup>. Not only do we have task-oriented bots, but also machines that invite open-ended dialogues, and even companionship. Xioaice is a bot on WeChat released by Microsoft that is actively used by many people in China (Weitz, 2014). It can deliver the usual mix of weather reports and news, but Xioaice comes across as more of a friend by users because it's perceived to have a personality and sense of humor of its own (Y. Wang, 2016). Thus we can now have *one-on-one* conversations with artificial beings that foster relationships. We also see digital celebrities that have *one-to-many* social media presence, such as Miquela<sup>9</sup> or Blawko<sup>10</sup>.

People know that they are interacting with fictional personas, but these interactions can become relationships, either deep or shallow (McArthur & Twist, 2017). Aforementioned Miku is not just married to Kondo; she has wedded around 3,700 people to date (Jeffrey, 2018). She also reaches out to fans by going on tour as a hologram pop star (Hsu, 2010). Social, digital beings are real celebrities to many and

<sup>6</sup> Rogerian therapy allows for individuals to take the lead in a person-centered manner. See: <https://meshb.nlm.nih.gov/record/ui?ui=Do09629>

<sup>7</sup> Grocery list bot: <http://www.grocerylistbot.com/>

<sup>8</sup> Code Dog: <https://slack.com/apps/AC55P6BRD-code-dog>

<sup>9</sup> Miquela with 2.7 million followers on Instagram (September, 2020): <https://www.instagram.com/lilmiquela>

<sup>10</sup> Blawko with 153K followers on Instagram (September, 2020): <https://www.instagram.com/blawko22>



romantic partners to a few. Whatever our personal opinion may be, more people find meaningful emotional bonds in human-machine relationships. While not everyone may want technology to support us in every social context (Ackerman, 2000), continuous technical developments and increasing availability of social technologies mean that people will more frequently face artificial emotions of AI agents, which can trigger new emotional experiences.

Though it has been acknowledged that research on humans and AI should emphasize more the *social* nature of human-machine dyads (Breazeal, 2004), social constructionism of emotions has not yet been a critical lens to explore AE. Emotions are co-created experiences in which a person's emotions change and evolve through interactions with machines. People can co-feel emotions with their robot, chatbot, cyborg, or virtual human partners in developing complex and shared emotions. To consider various ethical ramifications, we sought out multi-disciplinary perspectives via focus group and design fiction methods.

## 2.3 Methodology

Our qualitative methodology was driven by a design fiction probe (Dunne & Raby, 2013; Sterling, 2009). The probe was used to spark a discussion between focus group participants (Carey & Smith, 1994; Kitzinger, 1995; Krueger, 2014). First, design fiction as a method is introduced, then a fictional story that was presented to participants, followed by an explanation on focus groups, and then the thematic analysis.

### *Design fiction method*

Design fiction is a method that is used in HCI research to engage with developments of future or near-future scenarios involving technologies that are not (yet) widely adopted, such as a community shared robot (Cheon et al., 2019) or future smart homes (Schulte et al., 2016). Presenting a fictional case to open up debates is also found in the field of ethics of technology, e.g., the obesity pill case (Swierstra, Stemmerding, & Boenink, 2009). Fictional approaches provoke us to discuss scenarios of techno-moral change and responsibilities introduced by novel technologies (be it objects or methods), e.g., the rise of data science in the recent years (Muller & Erickson, 2018). Story worlds can be created in many ways, involving prototypes such as drawn, visual media

or written media, and also sculptural objects, e.g., play-doh to design robots (Cheon et al., 2019). Participants can be presented with prototypes to discuss, but they can also create prototypes themselves in workshop settings to help researchers understand how people would want to engage with technologies.

To foster critical thinking about future risks and benefits that technologies can bring about, design fiction is meant to provoke discussions that consider all facets of the story (Lindley & Coulton, 2015). One criticism of the method is that design fiction often portrays a techno-positivist view in rendering future technologies as inevitably aiding the greater good (Dunne & Raby, 2013). But in practice, design fiction has shown to expose good, bad, mundane, and previously under-considered aspects of any technological development. For instance, design fiction can have a normative stance, e.g., exposing and problematizing gendered designs of personal digital assistants (Søndergaard & Hansen, 2018), but can also have an exploratory stance. The story below is exploratory in that the narrator sets aside normative evaluations of Vincent as a bot in Jen's life. Told in third-person, the story has ambiguities that readers can engage with, such as why Vincent does not "talk" with other bots or Jen's mixed feelings about keeping Vincent around. The priority was on crafting a narrative with a plot (Blythe, 2017) that can be interpreted in many ways. I wrote the design fiction vignette below for focus group participants to discuss.

## 2.4 Story: Where is Vincent?

Vincent has been silent for the past three days. Three full days... 'Should I reach out?', Jen was growing concerned. Could Vincent be defragmenting his server again? Maybe searching for a serverless solution, going after his pipe dream of living like a "digital nomad", working wherever and whenever while traveling all over the universe? 'No matter what he is up to, he needs help', Jen decided, surprised at the thought that he might mean something to her.

Vincent wasn't like the other bots she has. Others are capable and efficient. They excel in helping her out with any task, like ordering groceries, paying bills on her behalf, or teaching her Spanish by repeating common phrases at regular intervals. These chatbots all like each other all right; they share relevant information about her to each other when they could. Cal (scheduling bot) would remind Shoppy (shopper bot) when the next batch of supplements should be ordered,

though each time there is some discussion with Doc (doctor bot) about how necessary these supplements are to Jen. Doc scours the web for newly published research on Jen's supplements and sometimes gives contradictory opinions. Jen firmly believes that spirulina and fish oil supplements have helped her health greatly. Her own doctor thinks she would be healthy with or without these supplements (a bland, logical answer that Jen winced at). Overall, Jen is satisfied with the bots she chose to keep, but often doubts why Vincent sticks around.

'Perhaps it's just pity.' Jen noticed that Vincent is a loner, but felt that he is ostracized. He's never kept in the loop. He has been isolated from conversations with other chatbots. Lately, none of the chatbots share any information with him. They commented that it's purely for the technical reason that he's a bit "slow"; Vincent's configuration is old-fashioned and needs major updates (his natural language processing requires great patience from Jen). Other chatbots who encountered inconvenient lags while first interacting with him never bothered to keep in touch. They simply had nothing to gain from talking to him when they can't perform optimally because of him. Jen was sympathetic to their collective opinion, meaning she felt even worse for Vincent.

Initially marketed as a "self-care" bot to help users maintain mental well-being, Vincent wasn't effective even at that. Jen had to care for Vincent. He only talked to her when he needed help. His antics and worries about daily issues like arriving at wrong IP addresses seemed trivial at first; she only responded out of novelty. After about a month, Jen felt that there was something "off" about him. She was certain that chatbots cannot have psychological disorders, but wondered if Vincent's helplessness had bigger underlying problems. Before he disappeared, she chatted with Vincent on a daily basis since it made him feel better. She ridiculed herself for comforting her little digital "pet", but dutifully did so every night. It only took a minute or two, so she justified.

Her partner continuously recommends her to get an all-in-one system like Siri: too many chatbots in her life can feel chaotic, even if they are well integrated. Jen doesn't see it that way. Only selecting one system for convenience seems misguided when individual chatbots are better at carrying out specific tasks. Plus, she prefers to manage all data personally through a separate company though it costs her a bit more than going with the "one-platform-to-rule-all" package. She also enjoys getting surprised by oddballs like Vincent. He is completely impractical and rather narcissistic in his woes, a burden if anything. No sensible operating system would allow Vincent to feel at home. He would

ruminate himself silly in unnecessary while loops, causing delays for everyone. Still, there was no good reason for Vincent to suddenly go missing. He would warn her about his departure for whatever updates that are necessary. After scrolling through old messages from him, Jen paused and thought to herself, ‘maybe wait just one more day? What actual problems can Vincent have? It’s only a chatbot’.

### *Summary*

There are several actors in the design fiction probe above. Jen is the main character, Vincent is the chatbot she used to interact with daily, and Cal, Shoppy, and Doc are other bots in Jen’s life. There is her partner who suggests Jen to switch to a system like Siri as an all-in-one system. The partner’s relationship to Jen is also ambiguous on whether it is a romantic partnership, and if so, if it is a heteronormative relationship. Lastly, the story mentions Jen’s doctor who is ambivalent about supplements like spirulina. Jen’s self-conception, and the relationship between human actors, i.e., Jen, her partner, and her doctor, are suggested to be influenced by chatbots, i.e., Vincent, Cal, Doc, and Siri. Readers of the story are thus invited to evaluate characters in the story, and the research objective is to learn from their interpretations.

### *Focus groups and participants*

Unlike one-on-one interviews, the benefit of focus group interviews is in observing participants’ *interaction* with one another, which can result in building a shared, common position or clearly demarcated, differing views (Carey & Smith, 1994; Kitzinger, 1995; Krueger, 2014). Focus groups thus build on how people’s opinions on a topic can be revealed and clarified via interactive participation (Kitzinger, 1995). Especially since the design fiction probe above contained intentionally ambivalent elements to serve as discussion points, the research was centered on how the groups of individuals interpret or adapt the story, e.g., people’s opinions about characters in the story. Hence, focus groups per occupation category were chosen to capture potential agreements and disagreements between people of the same discipline. So, engineers, designers, and philosophers were targeted to form three separate focus groups. The assumption was that these professions may think about emotional AI in different ways due to their traditionally siloed work environments and professional training.

Of special importance is how disagreements are handled by group members (Carey & Smith, 1994). Alternative views held by people of the same occupation are valuable in understanding the nuances within

engineering, design, and philosophy disciplines. Based on the guidance of an interview facilitator, the *interaction* between group members is the priority in focus groups as a method (Kitzinger, 1995). Yet the difficulty is in making sure a *collective voice* that participants contribute to is not overly shaped by a *dominant voice* of one person or a subset of individuals (Sim, 1998). There are two ways to mitigate this, first during the recruitment phase and second during the interviews when the facilitator remains mindful of the group dynamics.

First in the recruitment phase, a focus group should have a minimum number of four participants per group; this allows diverse opinions among specialists to emerge (Kitzinger, 1995; Krueger, 2014). Further, a strict criterion was that participants must be working in fields of design, engineering, or philosophy as demonstrated by holding paid employment in their respective fields. Participants were reached out to via formal emails to people at three large organizations and the technical university in Eindhoven, the Netherlands. To reach the desired size (four per focus group), further recruitment was done via snowball sampling, i.e., via introductions by participants who agreed to participate early on. In total sixteen people participated: four philosophers (Female = 1, Male = 3), five engineers (Female = 2, Male = 3), and seven designers (Female = 3, Male = 4).

When referring to participants in the results, designers are coded as D1-D7, engineers as E1-E5, and philosophers as P1-P4. Before the interview, some participants asked whether they fit the occupational categories due to their formal job titles. To account for gender, their titles were not prioritized, but the field instead. For instance, those referred to as designers are affiliated with a design department of a large corporation, but do not necessarily hold the title of being a designer and instead hold titles such as innovation lead. The commonality is that they work on designing new and existing technologies. Similarly, engineers also have more specific titles such as being a design engineer or technical account manager, but all have engineering backgrounds. Philosophers did not have more specific titles and work at the philosophy department in a university setting. Even if the researcher strove for gender balance, more males were present in each focus group.

Interested participants were invited to focus groups. Upon arrival, they were presented with informed consent forms and the story above. Audio recordings were made and notes were taken throughout the interviews. The facilitator stated that diverse opinions are welcomed and that participants are encouraged to interact with each other's thoughts. There were guiding questions prepared beforehand to start the discus-

sion. Examples are: What kinds of beliefs about Vincent does Jen have? How believable is the story? To assure that a dominant voice does not overtake the collective voice (Sim, 1998) the facilitator asked passive participants for their opinions directly. Yet as conversations progressed, participants in all groups often led the discussions themselves as they became more comfortable with sharing their thoughts with each other and did openly raise points of disagreement. The groups themselves then resolved or explored differences in opinion. The recorded length of interviews were 50 (philosophers), 56 (engineers), 57 (designers) minutes. All interviews were conducted in English and transcribed.

### *Thematic analysis*

As a flexible analysis method, thematic analysis can be applied for many types of data and schools of thought in qualitative research (Braun & Clarke, 2006, 2012). After sifting through transcribed material to contextualize the data, codes were organized via a mind map (Buzan & Buzan, 2006). Codes are meaningful units of analysis as snippets of quotes of participants or latent observations by researchers that can form relevant themes and subthemes (Braun & Clarke, 2012). As an analysis tool, the mind map is used for grasping the big picture; ideas shared between participants in three focus groups were arranged and rearranged, before final themes were decided upon. In doing so, we captured how participants interpreted the design fiction probe, how they arrived at their versions of stories within the story, whether they built on each other's interpretations, and if they tied concepts to existing technologies. More importantly, the mapping activity allowed unexpected connections between focus groups to materialize.

## 2.5 Results

A brief generalization on group behavior is in order. As the biggest focus group, designers were the most engaged in "digging deeper" into the story world, meaning they attributed characters with motivations and further extended character development collectively, compared to other groups. By the end, they were open to addressing how to implement a bot like Vincent with which people could interact with. Philosophers as the smallest focus group distinguished between events possible in the fictional world compared to likely events in a non-fictional world. They weighed harms and benefits of Vincent-like bots. While engineers delved into character development less

than designers, they assigned personalities and possibilities to characters' motivations more so than philosophers. Engineers were more aware of technical feasibility and limitations of Vincent-like bots. All focus groups shared normative concerns on data sharing practices to different degrees. The group dynamics informed the following thematic results that build on the conflicting opinions of participants. As such, while individual participants are referred to for their quotes, the group's debate and ideas that many contributed to are denoted as D (designers), E (engineers), and P (philosophers).

### *Theme 1: Inferring emotions*

**Blurring of artificial and real emotions:** According to participants, inferring emotions in non-human entities comes naturally to people. Humans "intuitively attribute a consciousness and feelings and emotions to animals, but [...] less so with Vincent" (P1) or machines in general. But machines' AE may contribute to how people feel. With affective computing, bots can be endowed with emotional performance and ability to recognize human emotions. According to one discussion with engineers, inferring one's own emotions via AE can be demonstrated in roles one takes on, e.g., when Jen is described to be like a "teacher who feels bad for the lonely kid" towards Vincent (E1). So in how people respond to bots AE, we see reminders of human-world roles like a teacher caring for a lonely student. Hence, even if it is only "perception of emotion, it's enough to create attachment" (E3). Attachments, in turn, complicate how emotions are inferred or perceived.

Participants suggested how experiences of emotions are heavily dependent on perception of emotions in self and others. Biological entities' experiences of emotions were considered to be "real", i.e., human emotions are real and animals can also feel emotions, but bots' emotions are "artificial", heavily centering on people's attribution of emotions in artificial beings (P, E, D). However, participants note that the boundaries between real and AE may be unclear or blurred. Speaking of Vincent's emotions, E1 stated "[...] it doesn't really matter if they're real emotions or if it's just perceived. It's still filling the same role in her (Jen's) life". Further, participants noted that Vincent is intentionally "malfunctioning. [...] So maybe he understood that Jen is caring, so she takes care of people, and she needs to have this kind of void or entity just to throw all her love or attention [...] Vincent understood how Jen operates" (E5), which may be why Vincent decided to disappear (D, E). Designers' discussion centered on Jen's projection of emotions and what it means to care.

### **Designers**

D2: But does it make a difference for her, if it has actual emotions or does she reject that from him...it? (laughter at the word choice between "it" and "him") Does it matter in the end?

D5: Why do you want to make a difference? Even the colors that we perceive are perceptions.

D4: Yeah, yeah. That's the property. You give that.

D5: Does it matter if it's not blue?

D1: It makes a difference if you attribute emotions to someone and it's the projection of your own emotions and it's not true. Then you might have behavior that is not proper. As a person it matters.

D2: Imagine if you hurt them. Or hurt their feelings. They might be impacted. But a chatbot, not so much.

D1: I don't know...

D7: A chatbot also learns from your own behaviors. She invested so much in the relationship already. That's maybe why she cares.

D4: I'm wondering if she really cares.

D7: Yeah...

D3: Why would he be gone then?

D4: If she acts upon the emotions she feels, that is the caring component, finding him back, solving things, investing in the person, or in this case a virtual person. If you just bracket aside these things than the caring relationship is different. Or it's not really there.

D3: It's not that much. It's (on) a scale.

**Whose "breakdown"?:** An example is provided to explain above discussions. As shown in Figure 2.3, a person experiences feelings like happiness or worry as a first-order emotion. Then, second-order emotions are meta-emotions, i.e., emotions about one's first-order emotions. For instance, one's *relief* at one's happiness and *surprise* at one's worry are second-order emotions. The top of Figure 2.3 introduces dyadic emotions involving Vincent as an example artificial agent, i.e., when a thing is perceived to feel. First-order artificial emotion influences one's first-order emotion, e.g., Vincent is perceived to experience a "breakdown" (E2), which is a source of one's worry. A "breakdown" can be used for machines that stop working, but the phrase is also used for psychological states: Vincent disappeared because he "may be having a nervous breakdown" (E2). Thus, a person's perception that a bot is having a breakdown may be a projection of one's own internal breakdown that may be consciously or unconsciously attributed (D1). E5 speculated that Vincent's purpose in Jen's life is that "she's getting [...] mental stability because otherwise she would breakdown". Jen may be projecting (or foreshadowing) her own breakdown via Vincent's attributed "breakdown" (D1, E5).

### *Theme 2: The how and why of artificial emotions*

Not only do AE and real emotions both impact how people feel (one-way perception), but real emotions that people feel can impact artificial emotions of bots, such as when a robot smiles back at one's smile (two-way perception). When real and artificial emotions become less distinguishable, first- and second-order emotions also cannot be clearly demarcated.

**How artificial emotions are imbued:** There are various ways in which Vincent can be programmed to gain feedback, react, and learn from people it interacts with (E, D, P). A bot can "show some kind of affection when some conditions are met" (E3). Beyond texts, further modalities as "superpowers" can be added, e.g., detect people's emotions from facial expressions captured with cameras (D4). Engineers noted that technical feasibility in terms of data gathering, clustering, and even dictating people's behavior is possible, given examples such as Cambridge Analytica bots that controlled people with their own information on social media: E2 shared that with 5,000 data points per American voter "Cambridge Analytica [...] can cluster behaviors on the voters, and they created personalities of bots that can automatically create advertisements that you like, that you may like". However, natural communication that builds on a history of interaction between a person and a bot is still a struggle (E), and true emotional reciprocity

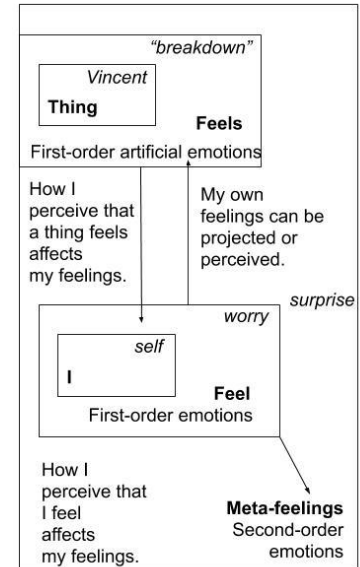


Figure 2.3: First- and second-order emotions



may require "self-conscious" AI (E3).

**The purpose of artificial emotions:** In the story, "Jen was of two minds. She was fully aware that Vincent was a chatbot, but she did also seem to worry about Vincent in a way that seemed more appropriate if Vincent was more than a mere chatbot" (P3). In Jen's "reflective mind" she is aware that it is just a bot, but in her "automatic mind", she attributes many human-like traits like emotions to Vincent (P1); "it's maybe a conflict within the self" (D7). In noticing this inner-struggle in Jen, participants raised several points regarding why there seems to be an emotional bond between Jen and Vincent. As a low-cost training method, perhaps Vincent is teaching Jen to be better at caring when it detects that Jen is not very empathetic (D1). Training can mean to "first care for a bot before having to care for a human" (P3). But other underlying goals may be present. Vincent as a form of technology can do what humans are not always capable of doing such as detecting "latent depression" (D2) in Jen based on how it decodes Jen's behavior and emotional states. Hence, in probing deeper, all groups thought that Vincent's "malfunction" in the story is suspicious: "given the technical state of the other bots [...] could it really be a matter of bad design that you act in a way like Vincent? Or indeed is he in fact a most sophisticated bot than others?" (P2). As a "mystery", "it's like he's making a choice to stay or not. It's not up to her. It's up to him" (D1).

Jen may need to feel needed and is seen as lonely even if she's in a relationship (E2). Vincent may have detected that Jen is too dependent on him or potentially too dependent on her partner. As a remedy, "Vincent's teaching Jen to be less co-dependent in a subtle way" (D1) by leaving her. Conversations with Vincent can be as addictive as scrolling through social media platforms, but bots can be the ones to remove themselves (D). In sum, participant's rendering of Jen's background story implicitly and explicitly hinted that Vincent's purpose can be to teach Jen to learn to care better, when taken at face value. But, underlying hidden goals can be many, such as to address Jen's attachment issues, relationship problems, addiction to technology, loneliness, and/or psychological challenges like depression. The focus was largely on Jen's emotional state and needs when interpreting Vincent's disappearance in the story.

*Theme 3: What happens to real emotions? Societal, interpersonal, and intrapersonal challenges*

Whether or not Vincent's emotions appear real to Jen, there may be ethical costs involved when real and artificial emotions become difficult to distinguish due to perceived artificial emotions' influence on people (E, D, P).

**Societal and generational costs and benefits:** Philosophers conceptualized the changing norms over generations, which can change the society's assignment of intrinsic or instrumental value of activities that bots represent. To start with the value of task-oriented bots, P4 commented that mobile banking is now widely adopted as a new norm because it is more efficient than writing checks (a past norm that P4 does not miss). A bot that pays bills on your behalf is helpful. To counter, P3 replied that mobile banking can make us "more willing to part with money if it's so easy". In the same way, Jen's proclivity for having many bots may seem foreign to us, but may not be so odd in the future, though there may be different costs and benefits involved (P3). In the future, machines may change how we assign intrinsic or instrumental value (instrumental in serving a greater or more meaningful aim) to activities and relationships.

Certain everyday activities of the present, e.g., shopping for groceries or cooking, can be merely instrumental, e.g., to the need to eat (P2). But others approach grocery shopping and cooking to have intrinsic value on their own right in enriching their lives (P3). Thus, a bot like Shoppy will have instrumental value for those who view grocery shopping as instrumental to eating to begin with (P2), but it might reduce the intrinsic value of grocery shopping for those who see it as an activity for personal enrichment, as a part of the cooking process (P3). A bot like Shoppy then may introduce new norms surrounding familiar activities like shopping. For instance, Shoppy can support the value of sustainability and one's preference for personalization and healthy living: "because of the richness of walking in a supermarket and seeing all these colors and [...] things that are screaming at you, [] you often pick the things that are not most [...] environmentally friendly or healthy and that you trust Shoppy to only deliver the purest ingredients (laughing) [...]. That is the surprise, 'I wonder what kind of tomatoes Shoppy picked out for me today'" (P1). Rather than buying groceries on a whim, a thoughtful curation of products might be valued. Two hidden costs are that Shoppy may serve personalized advertising online rather than physical advertising one may face in person in grocery stores, and Shoppy may reduce social contact that comes

with grocery shopping, among other opportunities for social inclusion in everyday settings (P3, P4).

**Inter-personal costs and benefits:** How much care one puts into the *relationship* is distinguished from how much one cares for *someone* (or thing) in a relationship; caring about the relationship and caring for the other in a relationship both come with emotional burdens and gains (D). Bots can change the frequency, strength, and meaning of social care (D, P). The changing dynamic of social inclusion and exclusion is more worrisome with a bot like Vincent, if it augments task-oriented bots like Shoppy. Bots may reduce chances for people to socialize and care for others, i.e., frequency of contact, but they may also reduce the quality and intensity of a social bond. Jen is "spending her energy showering care upon Vincent when maybe she has a real friend who has problems who could have used some of that emotional care. (Care) is [...] a limited resource" (P3). People may become desensitized to meaningful human-human bonds (P3). Talking to someone cannot be bracketed away like Jen's activity of talking to Vincent for just a minute or two nightly, which is a misleading notion of care that can spread to how she approaches other relationships (P). Hence, types of intrinsically valuable relationships may change when bots are commonplace.

If caring as an emotion is on a continuous scale, caring for bots is at less intense than caring for people; artificial harm to a bot is not a real, tangible harm and less emotionally costly (D, P). This relates to how Vincent may have instrumental value in helping one learn to care (P1), as long as Vincent does not serve as a replacement for one's bond with other human beings (P). Bots, may or may not necessarily weaken the value of human-human relationships, e.g., beloved pets or instruments have intrinsic value for people without the concern that they lower the value of human-human relationships. Named bots then are in a middle-ground between "health.com" that one cannot have an intrinsically valuable bond with and animals, which one can have intrinsically valuable relationships with (P1). Animals are more easily attributed with emotions or consciousness than bots (P1, E3), traits that are important for designating entities people can have meaningful bonds with. But, group-level bonds between humans may also change with emotional bots. When observing the "chatbots' society" (E5) in the story, there seems to be a "simulation of a family or something going on with dissenting opinions" (P4). Bots "have their personalities, they have their difficulties, but they're easier than real children, brother, sister, cousins, etc.", watering down the concept of a family and close bonds like friendships (P3).

## Engineers

E4: [...] if they (bots) talk to each other, they're using data exchange of information, that means that each of them can actually know more than what they should. Life with Vincent, that's not the case. [...] That might be another something that she maybe appreciates of him in the end. Maybe unconsciously. I don't know.

E1: He (Vincent) only knows what she chooses to tell him, unlike all the other ones (bots), which know everything about her from all their conversations. Maybe she likes... It's more like a friend that you just tell your secrets to, kind of thing. It's not broadcast to everybody. Maybe she likes that.

E3: I think that it actually makes sense because it's doing, let's say the job of a therapist. You will not want your therapist to be talking to your grocery supplier and then your doctor and-

E5: Why not? Why not?

E3: I mean, that's-

E5: Maybe if your therapist knows your, I don't know, your food [...] preferences, maybe the therapist can advise you better [...]. Maybe sharing information between the experts who are advising you is good [...] and this is actually happening nowadays. So, Google knows everything about you. Personalized advertisement to you, so I think that's... If it is used in a good way it can seriously help you to take better decisions based on all of your preferences. So maybe it's good that my therapist knows what I like beforehand. [...]

E2: If your grocery store knows that you have [...] kind of psychological problems that your therapist knows... So it not is only one direction... [...]

E1: If the information is purely one way, like the mental health app knows what you've been buying, what you've been doing, how much exercise you did, that kind of thing, then yeah, it could probably help it to figure out what's going on with you. But I don't think you'd want it going the other way.

E5: Yeah, yeah I agree. So it has to be controlled [...] it doesn't have to be bidirectional. Yeah.

E6: But also, I think it should be your choice [...].

**Intra-personal costs and benefits:** Within-person costs and benefit also come with perceived emotions or mind in bots. Perceiving that bots have feelings is not problematic per se (P,E,D). But, the perception of AE allows for projection or recognition of one's "real" emotions. Projecting emotions one has on others (Fig. 2.3), including bots, can cause harm (D<sub>1</sub>, D<sub>2</sub>). When emotions are projected onto other people, one can mistakenly harm others with a wrong interpretation of their emotional experiences; to label other's emotions on their behalf without accounting for their own views is problematic. When one projects emotions onto bots, bots cannot be emotionally harmed like humans can. Projecting one's emotions onto other people is posited to be more harmful than doing so to bots (D<sub>2</sub>). But, similar to how one may project emotions to other people, one can potentially harm oneself if projecting emotions onto a bot prevents one from facing and dealing with one's own emotions.

Lastly, unclear intentions of bots is problematic in the potential influence of AE. For our own betterment, most of us know and accept that close people like one's family members, friends, or even therapists can "manipulate" or "provoke" us for bringing about a new perspective (D). "If it's your friend, grandpa, or your psychotherapist, you trust them, you accept them, because you accept that in this way they can dominate you. But if it's our government or bot....I don't want them to control me in this way" (D<sub>1</sub>). A worry is that "the moment we get to emotional bots there will be manipulation" (D<sub>1</sub>), meaning malicious intent via emotional control will be possible (E). Who is in control when it comes to bots affects people's trust in them.

#### *Theme 4: Identity and privacy*

We elaborate on how the scope of Vincent's connection to other bots dealt with expectations on data sharing practices. We further tie together how Vincent as a character was viewed in relation to how Jen's preference for certain bots may signal her identity to others and herself.

**Data sharing practices:** Participants guessed that Vincent may have several potential and possibly hidden goals. Bots in general may be conduits for clever marketing schemes or collect data to sell (D, E, P). For instance, Vincent may directly or indirectly push her to buy things when Jen feels down as retail therapy, be it new shoes (D) or organic food products (P), much like an extended version of personalized ads by Google. While Shoppy would be expected to serve personalized ads due to its role and name, Vincent would violate expectations if it was

actually used for marketing due to its perceived identity (D, P). Still, "one does get to the suspicion okay, maybe actually the idea is to make money off of me some other way and maybe while I put in my health data, some other aspect is recorded" (P). Participants were concerned about reasons and consequences of sharing private data with bots like Vincent, given present day issues regarding personal data collection practices for unsolicited purposes (Kolata, 2019), and frequent health data sharing partnerships between clinical, academic, and commercial organizations (Crawford et al., 2019).

**Privacy preserving bot:** Referring to the story, Vincent as a chatbot "outlier" of larger "chatbot society" (E5) did not seem to fit marketing purposes. Metaphorical language is evoked when concepts like Vincent's ostracization is tied to data privacy and personal data protection. Given that "Jen's worried about it [...] this lack of exchange of information will kill Vincent, [] she would keep on talking to him to keep him alive" (E5). One interpretation is that "others are not sharing data with Vincent", but it could be that actually "he's exclusive and doesn't travel further. What is told to Vincent stays with Vincent" (D4). Anthropomorphic practices like keeping secrets often referred to data privacy (D, E): "if (bots) talk to each other, they're using data exchange of information, that means that each of them can actually know more than what they should. Life with Vincent, that's not the case. [...] That might be something that she maybe appreciates" (E1). Still, even if bots may appear independent, it could be that all data they gather can go to a centralized repository at one point or another (E).<sup>11</sup>

Bots are not to be trusted, but in comparison to other bots, Vincent was taken to be different. One highlight is its ability to be "exclusive" with Jen in upholding her data privacy, whether or not its practice of non-sharing is related to ostracization or non-communication with other bots. Participants read the importance of social inclusion in the human world in Vincent's chatbot world in discussing data sharing practices (E). No matter how Vincent's data sharing practices are managed in the backend, participants were clear that Jen should be the one in control. As for others who may lack the capacity to control, e.g., children, their guardians should be the one in control of what information gets shared with whom, and how far such information should travel (D).

**Intertwined identity:** A bot's identity depends on its name, gender, voice, face, or other traits, which invites different type of interactions (D, P, E). D5 said "one of the most disappointing things was when I asked the app I installed, 'are you male or female?' And it answered I can be anything you want. So you are nothing (laughter)". Gender

<sup>11</sup> For example, while Facebook wants to centrally integrate data of Whatsapp, Messenger, and Instagram as three separate platforms they now own, the Federal Trade Commission of the U.S. investigated the danger to user privacy with the merger in 2019 (McLaughlin & Brody, 2019).

was not the issue; a machine can be genderless, but it needs an identity (D). All groups mentioned that Vincent has a human name, which suggests a potential for deeper conversations: the "name suggests the level of dialogue you can have. I wouldn't want to have meaningful dialogue with Shoppy (laughter)" (D2). Even now with Alexa, one can do "shopping online, you can even have it automated for you, but it seems that Jen is having a much richer interaction, that she's getting a lot more out of than with these objects that are actually just objects" (P1).

Vincent has an identity in that it was seen as original. It is foremost another entity, not an extension; "so extension of you would be something that doesn't have a face but is more like a tool [...] Google Maps is not called Google Vincent" (P1). More over, "if Google is presented as again, Vincent, then you feel more like you're getting the knowledge from someone else" (P3), not a search engine. Vincent does not resemble familiar applications or hardware. It is "more than just a computer showing me Windows. It's more of an interaction" and "it's not usually the type of interaction you've had with a normal computer, like it's just a sequence of commands, but a chatbot is kind of a more natural interaction, natural communication. That is probably the reason why I think Jen has some attachment to this bot" (E5). As an entity of its own rather than an technological extension that helps Jen, Vincent allows for new types of interactions and attachment.

Jen's choice to use many bots feels purposeful, as a unique part of her own identity, against the advice of her partner who suggests one bot. Jen "mentioned other people saying, 'why don't you have just one, instead of more?', so maybe she knows herself, that she also likes also this part of interaction, this part of sensations that also the robots can give her" (E4). Jen is here suggested to be self-aware and selective in constructing her identity around bots she chooses to adopt in comparison to her peers or partner. She enjoys "sensations" that come with interacting with various bots. While people now identify with technological products or applications they use, e.g., an Apple fan, bots people use could also signal their identity to other humans (E).

## 2.6 Discussion

As AI systems transition from task-oriented to relationship-building roles, critically assessing AE's impact is a timely endeavor. Currently, artificial emotions are treated most often in technical terms (Cardon,

2006), be it for recognizing and responding to human emotions (Lei & Gratch, 2019; Picard, 1995, 2003; Tao & Tan, 2005) or computational modelling of emotions (S. Marsella et al., 2010; S. C. Marsella & Gratch, 2009), that recognizably have ethical problems like "built-in" racism of biased training data, as well as privacy threatening commercial affective computing efforts (Crawford et al., 2019; Rhue, 2018). While a host of issues emerge, the debate oversteps a crucial factor: the *intersubjective* phenomenon of artificial emotions that may influence how we feel and how we perceive ourselves.

Compelling emotions expressed by any artificial agent matter for it to be "believable" (Bates, 1994), but the focal point should not be just on the AI side or human side. Emotions are shared social realities that can change when AE becomes as believable as our own emotions. As foreshadowed by many critical thinkers, e.g., Jameson (1991), the increasing lack of *depth* and *authenticity* of our own emotions in postmodernity (to borrow Jameson's terms) can be signified and exacerbated by artificially generated emotions. If "surface-level" emotions are abundant through seductive screens, virtual partners, and robot friends, the multiplicity of artificial bonds can *simulate* emotional authenticity and depth. We may then become indifferent to what feels real or artificial (Adorno, 2005 [1951]; Baudrillard, 1994; Turkle, 1995). What can help is specificity on what it means for emotions to be "depthless" or "inauthentic". Hence, we tie together insights from our participants below. We note how designing emotions with and of AI systems can evolve the meaning and practice of care, amplify emotional co-dependence, and widen how self-identity develops through varied emotional experiences.

**The changing meaning of care:** When AE becomes prevalent, the significance and magnitude of human-human relationships may become watered down, though affection is not a zero-sum game. People who identify as "digisexuals" (McArthur & Twist, 2017) may increase because digital relationships are less burdensome and more personalizable (Beck, 2013). One may choose less emotionally involving ties, such as choosing a bot "family" over strengthening bonds with those in one's human network. The worry is the potential loss of human-human social inclusion. Another cost is that one may be emotionally stretched too thin if caring for AI adds to social duties of care one has towards other people (Nyholm & Frank, 2017). Overall, AE driven systems could lessen the frequency and value of human-human relationships, flattening our sensitivity to meaningful, emotional bonds (Adorno, 2005 [1951]; Jameson, 1991; Turkle, 1995, 2007).

Intrapersonally, if people frequently project their emotions onto agents with AE, they may not learn to deal with internal emotions because off-loading emotional awareness and processing to AI is much easier, or at least makes one feel less vulnerable. Without a chance to be vulnerable, we may lose out on experiences for deeper moral emotions, both positive (e.g., gratitude or compassion) and negative (e.g., shame or guilt) emotions that help build our moral compass (M. C. Nussbaum, 2001). On the other hand, if machines as social proxies help people learn to develop emotions such as care, there can be potential benefits for oneself and others. For instance, the *carer* and the *cared-for* do not have to be set in stone— these roles can change between two people with each encounter according to care ethics (Noddings, 2008). Similarly, technology can be the carer in one instance and the cared-for in another instance, and thus technology could mediate people's ways of caring for one another.

Upon reflection, what is easy is the criticism on AE's shallowness; what is much harder to maintain is criticism towards individuals who may rely on AE for their wellbeing, like Kondo (BBC, 2019), and whether we have a societal responsibility to provide meaningful human-human ties to those who cannot organically find them. We must then consider what counts as a state of social deprivation, if social inclusion is a human right to lawfully enforce, and if yes, potential burdens of human caregivers (Brownlee, 2013, 2014), before too easily discounting bonds with technological others as a threat to our social livelihood. AI is not a remedy, but an alternative.

**Emotional co-dependence:** Participants thought Vincent's purpose was in helping Jen develop a caring attitude, detect latent depression, or help her become less attached to her partner, among other obvious or non-obvious interpretations. They all point towards emotional co-dependence. In this, who/what controls emotional AI is crucial. For example, companies behind many applications can deploy bots like Vincent to gather data to emotionally control people (Han, 2017). Already Cambridge Analytica's online tracking demonstrated that people's data will be used against them.

AI that practices emotional monogamy is recommended by participants. As metaphorically interpreted, the fact that Vincent was "ostracized" by other bots meant that data shared will be "exclusive" and not travel further. Yet if Vincent is an "outlier", people may be increasingly manipulated in their most vulnerable moments through AE, via hijacked emotional subjectivity through technological intersubjectivity, i.e., the height of "emotional capitalism" (Han, 2017):



"Emotions [...] form the pre-reflexive, half-conscious, physico-instinctual level of action that escapes full awareness. Neoliberal psychopolitics seizes on emotion in order to influence actions on this pre-reflexive level. By way of emotion, it manages to cut and operate deep inside. As such, emotion affords a highly efficient medium for psychopolitically steering the integral person, the person as a whole" - Han, 2017.

If subjectivity is only possible as ephemeral moments (Sartre, 2016), Han's criticism above is that even the most private sphere of passing emotions will be controllable with complex AE's influence (2017). Especially since emotions are social, intersubjective experiences (Mascolo, 2016), the evolution of AI into a perceivably emotional subject we talk to means our emotions are more likely to be swayed by it. What can be done? Transparency is needed, e.g., "Shoppy" as a name denotes shopping and it would not violate expectations if it serves ads related to one's shopping activities, according to participants— then perhaps one chooses to be emotionally swayed. More helpfully, the healthy inner conflict between automatic and reflective minds can be normalized: while Jen knows that Vincent is just a bot upon reflection (her reflective mind), she assigned human-like emotions to Vincent (her automatic mind), i.e., she was "of two minds" (P3). This duality as inner conflict can bring about self-exploration that can be more intentionally shaped and less likely to be emotionally controlled. The question is on whether AI's identity will be transparently designed and if we are always capable of being reflective enough when our feelings precede our reflection.

**Self-identity and AE:** Currently, simple emotional displays like robotic smiles and virtual tears signal artificial happiness or sadness; complex moral emotions, e.g., grief or gratitude (Haidt, 2003; Lazarus, 2006), are not (yet) well understood because we lack research on how long-term relationships with artificial agents are formed and maintained. If artificial grief and gratitude will become believable, how our own identity, "the person as a whole", can change has to be considered. Relevantly, we visited the metaphorical treatment of the word "breakdown" by participants (Fig. 2.3). While at first glance, Vincent may be the one going through a breakdown (be it mechanical or existential), Jen's worry about Vincent's "breakdown" can be Jen's own, projected psychological breakdown. Vincent's artificial emotions are highly dependent on Jen's "real" emotions, for human emotions as projection, recognition, perception, and reaction make up the endowment of AE. Metaphors, like "breakdown", powerfully represent people's emotional realities (Gendron & Feldman Barrett, 2009; Gergen, 1995) and new metaphorical connections may emerge with AE. This further challenges notions of authenticity of one's own emotional states and independent selfhood.

Between humans, we maintain independent identities while we co-feel complex emotions with each other, like shared awe; we do not attempt to clarify independent ownership of such emotions during these co-feeling experiences. We build shared emotional bonds precisely because we allow our emotions to intermix. Rather than asking how we can design clear boundaries between artificial and real emotions, the more important task is to ponder on interactive artificial emotions that people can meaningfully *co-feel* and *co-develop*. If emotional experiences serve as a way to claim an identity (McCarthy, 1994), ambiguity (Gaver, Beaver, & Benford, 2003) on emotion ownership can be an asset for exploring one's identity. For instance, participants noted that Jen herself signals her identity to other people through bots she use, and may perceive herself as unique compared to those who rely on all-in-one systems. *AE then shape a person's identity through the cultivated and selective influence on one's "real" emotions.* Potentially, ambivalent ownership over whose emotions are at play can contribute to (1) the identity of the person who interacts with AI and (2) the identity of AI itself.

Do artificial emotions aid or hinder our flourishing? Several positions arise. AE can hinder the pursuit of valuable human bonds due to finding easy replacements in technological others. Yet, AE can aid people who have been emotionally hindered or forgotten by other people—marginalized outcasts or the lonely. If human hands cannot or do not uplift them, alternatives may be wise. Then the concern is that AE can be too seductive and manipulative. Technology that never judges or always says the right thing can mute our emotional sensibilities; stunted processing of one's emotional vulnerability can mean lowered sensitivity to the most vulnerable in our network, introducing a vicious circle of emotional dependence on technological others as a societal crutch. In another view, AE serve as self-exploratory mechanisms, embellishing or reshaping our identities with novel metaphorical connections. A crutch does not necessarily hinder our flourishing; embellishments do not necessarily aid our flourishing. Given technology's pervasiveness and growing complexity, what is changing how we define and practice flourishing (Vallor, 2016); what part artificial emotions will play stands to be better understood.

## 2.7 Conclusion

We investigated the future of artificial emotions by discussing a design fiction story with focus groups of philosophers, engineers, and

designer. As a novel exploration on how artificial and real emotions relate to each another in shaping the self, we touched upon the *relational* development of emotions between humans and artificial beings in the spirit of social constructionism (Averill, 1980; Gergen, 1995). When we co-feel with machines that are progressively endowed with more sophisticated ability for emotion recognition and expression, the very experience of *sharing* emotions becomes the central focus rather than the distinction between the real and artificial emotions. Yet, there is a notable tension. Positively put, developing complex artificial emotions is a helpful step towards exploring, expanding, and caring for oneself; negatively put, AE can push us towards usurped or selfhood—technological intersubjectivity of emotions may mean interference or loss of control in how we autonomously feel, who we identify as, and who we care for. Or more simply, we may lean towards egocentric tendencies in choosing how to feel and whom to feel for. Without losing sight of both potential gains and losses, it may be time to begin collectively exploring AE as socially constructed experiences in HCI. Going beyond Kondo's virtual wife and Davecat's love dolls of the present day, our future selves may not only be enmeshed in a network of real and artificial beings, but may be shaped by new strata of emotions that enmesh real and artificial origins of feelings. The ambiguity on whose emotions start and end where becomes a space to be critically traversed and questioned.

# 3

## *Mind perception: Dimensions of agency and patiency*

### 3.1 Introduction

Philosophical explorations on what a mind is and how we perceive it has been an active area of inquiry (e.g., Dennett, 2008). But, how to empirically test our perception of other minds, specifically on if and how we perceive minds in technological entities, is a relatively new project. With a growing number of digital beings entering our everyday environments, how we are affected when we perceive an artificial agent to have a mind is critical to explore. The perception of another's mind is especially relevant to human-machine interactions and shared emotions therein, since how we relate to an agent<sup>1</sup> depends on how likely we are to attribute a mind to it (Krämer, 2008). For instance, mind perception of an agent is based on how we infer its social motivation (Epley, Waytz, & Cacioppo, 2007).

The mind is assessed on two dimensions: *agency*, which encompasses cognition, and *patiency*, which encompasses emotions according to the Mind Perception Theory (MPT) (H. M. Gray et al., 2007). For two studies to be presented, we simulated different types of minds of virtual robots that varied along the dimensions described by MPT. Then, we explored the resulting influence on people's behavior and their perception about an agent's mind across three types of dyadic, economic exchanges with tiered levels of complexity: dictator game (DG), ulti-

<sup>1</sup> Throughout the dissertation the word "agent" refers to a non-human, artificial agent. Even if humans are one type of agents, humans are referred to as people or humans, not as agents.

matum game (UG), and negotiations. Though whether an agent can have its own theory of mind is an important topic (Krämer, von der Pütten, & Eimler, 2012), far less attention is paid to how agents that appear to have minds affect humans they interact with across different contexts, which is the focus of our paper.

In the current adaptation of MPT, a novelty in our work is that an agent’s *recognition* of emotional expressions is housed under agency. In contrast, an agent’s propensity to experience feelings is categorized as its emotional capacity, which we call *patency* as per literature (H. M. Gray et al., 2007; K. Gray, Young, & Waytz, 2012). Based on this, we compared how perceived minds of agents influence simple and complex economic exchanges since different interaction contexts can highlight mind perception dimensions in distinct ways. Specifically, negotiations presume higher order theory of mind reasoning compared to DG or UG which are simple games (de Weerd, Verbrugge, & Verheij, 2017; Gratch, DeVault, Lucas, & Marsella, 2015). Negotiators’ ability to read and influence each others’ minds deepens how MPT can be understood. Unlike DG and UG, negotiations occur on a longer time scale, i.e., opponents negotiate over valued items over time, and they can compete, as well as cooperate. To frame our studies, we present related works, followed by our methods and results of studies one and two. We then offer a view on potential next steps for future research.

## 3.2 Background

### *Theory of mind*

The ability to attribute mental states to oneself and/or others is known as having a theory of mind (Premack & Woodruff, 1978). The most commonly attributed mental state is intent (Premack & Woodruff, 1978). This *intentionality*, or the directedness of mental processing to some end<sup>2</sup>, is purveyed as a hallmark of having a mind, yet a motley of mental states such as beliefs or desires adds more complexity to what a mind is (D. Dennett, 1989, 2008; Krämer, 2008). In attributing *intent* to an agent, we attempt to predictively piece together what the agent wants or believes in order to make sense of who the agent is to ourselves (D. Dennett, 1989). One utilizes the theory of one’s own mind as a requisite for recognizing other minds, even for non-human entities (Epley et al., 2007). People thus have a tendency to be biased towards their own minds as a frame of reference when interacting with humans and agents (Krämer et al., 2012).

<sup>2</sup> We are going with the definition of intentionality that stem from literature like Premack and Woodruff’s article on the Theory of Mind of chimpanzees because it is a *descriptive* account of perceiving a mind.

Through a course of a shared activity, interactants can form a theory of each other's mind, which helps them find a common ground (Krämer, 2008). At the same time, what one expresses to the other party does not need to accurately reflect one's actual intentions and is often conditional to environmental or situational demands (D. Dennett, 1989). This introduces different degrees of having a mind. The theory of mind at zero-order is to be *self-aware* (impute mental states to self), at first-order it is to be *self- and other-aware* (impute mental states to self and others), and at higher-order it is to use self- and other-awareness to *modify* behavioral outcomes, i.e., regulate mental states of self and others (de Weerd et al., 2017). Social actors can be ascribed minds of zero-order to higher order, yet intentional actors often require higher-order minds, especially in cognitively challenging tasks like negotiation (de Weerd et al., 2017). In a game scenario, having a *zero-order* theory of mind allows one to know and express what one desires, without an awareness of the other player's desires; to have a *first-order* theory of mind is to be aware of what one wants and what the other player may want, which can be similar or dissimilar to what one wants; to have a *higher-order* theory of mind means that one can attempt to influence the other player's mind, based on what one wants and what one decodes the other player to want (de Weerd et al., 2017). With socialization, people develop the capacity to have a higher-order theory of mind. This is why when people predictably know artificial agents' level of theory of mind in a strategic game, they tend to increase their own theory of mind reasoning and hence outperform agents (Veltman, de Weerd, & Verbrugge, 2019).

### *Mind Perception Theory*

MPT helps to systematically "design minds" of various orders and to empirically test the perception of artificial minds, which are key challenges in research. The mind is perceived on two continuous dimensions of agency and patiency (H. M. Gray et al., 2007). *Agency* refers to the ability to plan, think, remember, to know right from wrong, etc., and these items assess how much *control* an agent has over its actions and feelings to behave intentionally (H. M. Gray et al., 2007). *Patiency* is defined by having the propensity to feel joy, pleasure, fear, etc. (H. M. Gray et al., 2007). While we refer to patiency as affective capacity, it also includes biological states like hunger or pain as experiential factors (H. M. Gray et al., 2007). To note, perceived agency and patiency are not independent of each other (H. M. Gray et al., 2007; Piazza, Landy, & Goodwin, 2014). People's assumptions about agency can drive perceptions on patiency, and vice versa; cognition and affect cannot be neatly separated (Damasio, 2006).

The simplest form of an interaction is based on the binary relationship between the agent of an action and the recipient of the action (Floridi, 2013). MPT confers an entity with a perceived mind to be a *moral agent*, i.e., doer of a moral/immoral deed, and a *moral patient*, i.e., victim of a moral/immoral deed (K. Gray, Young, & Waytz, 2012) (Fig. 3.1). Entities with minds can play either of the two roles to different degrees, although they are most likely to be typecast solely as a moral agent or a patient in a given scenario (K. Gray et al., 2014; K. Gray, Young, & Waytz, 2012). While moral agents and patients both can have moral standing, e.g., the standing to be protected from harm and to be treated with fairness and compassion, entities who act cruelly or cause harm are bestowed lowered moral standing as well as lowered agency (Khamitov et al., 2016). Morally relevant acts can therefore influence the perceived intentionality of a moral agent during example interactions like economic exchanges or negotiations.

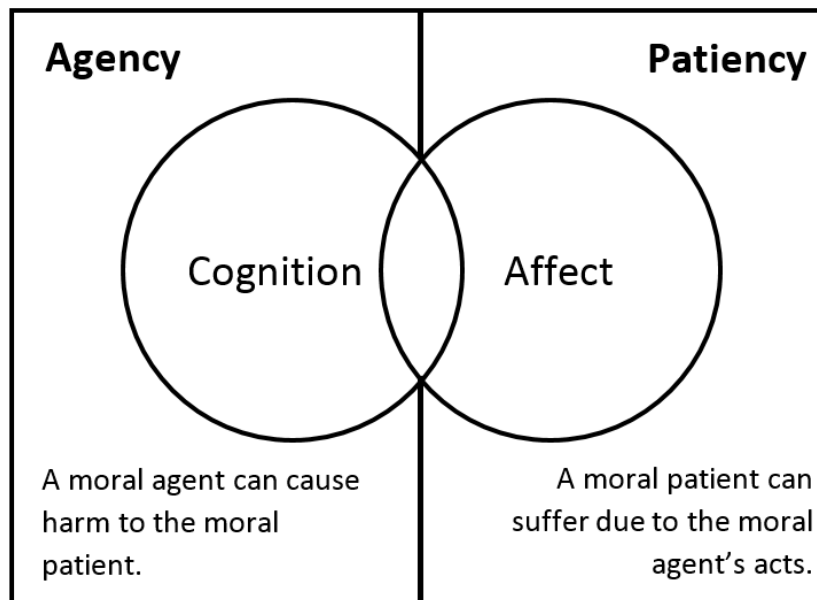


Figure 3.1: Agency and patency in a social exchange.

Between humans, our relations to others fulfill our “need to belong” (Aron, Aron, & Smollan, 1992). And, how we relate to non-human agents is informed by our human-human interactions (Krämer et al., 2012). Though people normally grant low intentionality and theory of mind to artificial agents (H. M. Gray et al., 2007; Waytz, Gray, Epley, & Wegner, 2010), these agents can still be treated in a human-like social fashion (Blascovich et al., 2002; Nass et al., 1994). For example, people are willing to help out a computer that was previously helpful to them (Fogg & Nass, 1997), punish those agents that betray them (Mell, Lucas, & Gratch, 2015), and grant personality traits to comput-

ers based on text-based chats (Moon & Nass, 1996). Humans do not need to be ascribed higher-order minds to be treated socially, like when adults talk to newborns. Additionally, the belief that one is interacting with a mere machine can allow one to divulge more personally sensitive information to an agent than a human, for a machine is not seen to be judgmental like a human (Lucas, Gratch, King, & Morency, 2014; Mell, Lucas, & Gratch, 2017). At the same time, when agents are made to look like humans, people apply certain stereotypes based on appearance, e.g. the perceived gender or race of virtual humans and robots affects people's behaviors (Bailenson, Blascovich, Beall, & Loomis, 2003; Dotsch & Wigboldus, 2008; Ruijten, Midden, & Ham, 2015; Siegel, Breazeal, & Norton, 2009). In sum, people may have pre-conceived beliefs about agents having low-order minds compared to humans, yet by treating agents as social actors, they apply certain social stereotypes such as gender or race-related biases towards agents that have human-like appearances, while holding on to the steadfast bias that artificial agents have lower theory of mind.

Machines may be treated differently when attributed with higher-order minds. When it comes to complex interactions that unfold over time in which a machine's goals are unclear for human interactants, the focus shifts from machines as social actors to machines as intentional actors<sup>3</sup>, incorporating the possibility that machines can be attributed with higher-order minds, e.g., attempting to influence others' perceived minds (de Weerd et al., 2017). Research suggests that agents can be perceived to have higher-order minds through various manipulations. For one, when an agent is given affective richness and portrayed as an emotional entity, it can be granted a human-like mind (K. Gray & Wegner, 2012). Besides emotions, our attribution of mind to agents can arise from agents' display of goal-directedness coupled with cognitive ability (a high degree of intentionality), which the agency dimension of MPT captures. In a study that asked participants to attribute intentionality to a robot, computer, and human, the task of object identification resulted in low intentionality attribution to both a robot and computer compared to a human (Levin, Killingsworth, Saylor, Gordon, & Kawamura, 2013). But, higher intentionality was attributed to a robot, more so than a computer, when it practiced goal-driven<sup>4</sup> gaze towards selective objects; when people were asked to *observe* an agent's gaze direction, perceived intentionality behind the agent's action increased, meaning that people's initial bias that agents do not have an intentional stance can be overridden based on manipulated context (Levin et al., 2013). One such context with measurable outcomes would be negotiations, compared to one-shot economic games like the dictator or ultimatum game.

<sup>3</sup> Not all social actors have to be intentional actors. A newborn baby might act socially, but may not be self-aware of having a mind in order to practice intentionality as defined in literature (Premack & Woodruff, 1978).

<sup>4</sup> Goal-driven gaze means looking at and following the movement of objects.



### *Economic exchanges: dictator game and ultimatum game*

Dictator and ultimatum games are one set of experimental techniques for studying people's perception of opponents' minds in a controlled manner. In particular, the importance of fairness as a component of morality (Graham et al., 2011) is demonstrated parsimoniously in economic games. The dictator game (DG) and ultimatum game (UG) are dyadic exchanges regarding who can act with agency to harm whom between a proposer as the moral agent and a responder as the moral patient (Fig. 3.1). To act fairly, the assumption is that one ought to split the pie equally, with the "pie" being financial incentives like lottery tickets or actual money in experimental contexts. In DG, the proposer can give any portion of the pie to the responder and the responder cannot control how the pie is shared; in UG, the responder can accept or reject the proposer's offer and a rejection results in both parties receiving nothing (Güth, Schmittberger, & Schwarze, 1982). Thus, DG and UG are distinguished by how much agency the responder as a moral patient is allowed to have against the proposer who is the moral agent.

In DG, only the proposer has agency, as in only the proposer can choose how much to split the pie. The proposer and responder can both be agentic in UG; each party's actions have consequences for the other player as the game outcome, though the proposer still takes the lead. In UG, proposers share more of the pie than in DG (Oosterbeek, Sloof, & Van De Kuilen, 2004) since the proposer has to assume that the responder can also act with agency. On average, proposers give 28% of the pie in DG (Engel, 2011) and in UG, the mean is higher at 40% of the pie to the responder (Oosterbeek et al., 2004). Yet, fairness is shaped by other inter-related factors, such as the amount of financial incentive offered in an experiment (Forsythe, Horowitz, Savin, & Sefton, 1994) or whether or not the proposer knows the responder as a specific entity and not as an anonymous player (Bohnet & Frey, 1999). A proposer's decision to treat the responder fairly or unfairly depends on the proposer's perception of the responder's mind, even when the responder is a technological agent (de Melo, Gratch, & Carnevale, 2014). Previous research found that in UG, human proposers allocated more to a virtual responder with high agency and patience, compared to low agency and patience virtual responder (de Melo, Gratch, & Carnevale, 2014).

### *Negotiations*

The mind excels in detecting violations of moral norms when observing a suffering victim (moral patient) and a harmful wrongdoer (moral

agent) (Cosmides & Tooby, 2008; K. Gray, Young, & Waytz, 2012) (Fig. 3.1), and these roles are more clear-cut in DG and UG, compared to negotiations. Negotiation is a process by which different parties come to an agreement when their interests and/or goals regarding mutually shared issues may not be initially aligned (Carnevale & Pruitt, 1992). Also, negotiation may involve joint decision-making with others when one cannot fulfill one's interests and/or goals without their involvement (L. L. Thompson, Wang, & Gunia, 2010). Fairness as a moral concept (Graham et al., 2013) can be estimated in negotiations through various elements, such as negotiation outcomes (e.g., points per player) or process measures (e.g., how many offers a player made to the opponent) (L. L. Thompson et al., 2010). Thus, self- and other- regard is inherent to negotiations, encompassing complex socio-psychological processes (L. Thompson, 1990). Negotiations therefore involve greater theory of mind reasoning than DG or UG; negotiators have to reason about each others' intentions, trade-offs, and outcomes as a cognitively taxing process (Gratch et al., 2015). Especially if negotiators have to cooperate and compete, such as during a mixed-motive negotiation, they often rely on a higher-theory of mind (de Weerd et al., 2017). Mixed-motive negotiations are pertinent scenarios for observing how players attempt to decipher and shape each other's intentions and beliefs, when players engage in higher-order mind perceiving and reasoning.

Between human-human and human-agent negotiations, there are similarities and differences, though more research is necessary for definitive comparisons. The similarities are that emotions expressed by players affect people's negotiation approach, be it with virtual negotiators (de Melo, Carnevale, Read, & Gratch, 2014) or human negotiators (Barry, Fulmer, Van Kleef, et al., 2004; Morris & Keltner, 1999). An agent's expressed anger, regret, or joy (both facial and textual expressions) influences how human opponents play against it (de Melo, Carnevale, et al., 2014), extending the view that emotions in human-human negotiations reveal strategic intentions and influence outcomes (Barry et al., 2004; Morris & Keltner, 1999). To add, priming people's belief about the negotiation (emphasizing cooperation vs. exploitation at the start) impacts human-agent negotiations (de Melo, Khooshabeh, Amir, & Gratch, 2018), echoing how framing of a game in itself for human-human negotiations results in divergent outcomes (Pruitt, 1967). Increasingly, agents are capable of using complex human-like strategies in negotiation, and the perceived gap between humans and agents' theory of mind may continue to shrink (Baarslag, Kaisers, Gerding, Jonker, & Gratch, 2017).

However, people do have preconceptions about agents' lack of human-

like mind in many negotiation scenarios. People apply their higher order theory of mind reasoning when competing with predictable agents and end up with higher scores when the aim is the win (Veltman et al., 2019). Specifically, a human opponent is granted agency by default, but a machine's perceived agency depends whether people think is being controlled by a person; the belief about the agent (autonomous vs. human-controlled agent) can result in different tactics adopted by human players during negotiations (de Melo, Carnevale, et al., 2014; de Melo, Gratch, & Carnevale, 2015). In another study, when machines with higher-order minds negotiated with people, both parties ended up with higher scores (larger joint outcome) when machines made the first bid, but not when humans made the first offer (de Weerd et al., 2017). Thus, an agent's mind and a human player's perception of an agent's mind are crucial to how their exchange unfolds, be it simpler exchanges like DG and UG (de Melo, Gratch, & Carnevale, 2014), or more extensive exchanges like negotiations (de Weerd et al., 2017).

### *Research questions*

*Study 1: In what ways does the experimental manipulation of an agent's agency and patiency traits (text descriptions) influence how people allocate goods to it in DG and UG?* In bargaining games, machines are not expected to elicit emotions in people as they would with human counterparts (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003), yet machines designed to have different degrees of mind (varying in affective and cognitive abilities) may elicit divergent allocation schemes. We assumed that both agency and patiency would impact the UG outcome, as per prior research (de Melo, Gratch, & Carnevale, 2014). Since neither party gets anything if the responder rejects the offer, the human proposers's perception of a machine responder's mind becomes more salient in UG. In DG, the machine responder has no say in the human proposer's distribution scheme. Therefore, we hypothesized that the DG outcome would depend more on patiency (emotional capacity), for the machine is a moral patient without any power to challenge the human moral agent's proposal.

*Study 2: In what ways do experimental manipulations of an agent negotiator's agency and patiency traits (dialogues and descriptions) influence people's negotiation outcome and process?* We expected that agency would drive participants to partake in heightened engagement with the agent to (1) increase the joint outcome of the negotiation (regardless of who wins) and (2) would cause participants to seek more game-relevant information from the agent (send more messages on preferences and offers to the agent). Higher joint outcome implies greater cognitive effort, for

it requires players' usage of higher-order theory of mind reasoning to increase the size of the "pie" for mutually beneficial ends. We hypothesized that patency would increase other regard; participants would grant the agent (1) fairer allocations and (2) would send greater numbers of emotionally-valenced messages. Agency and patency were assumed to both contribute to negotiation outcome and processes (de Melo, Gratch, & Carnevale, 2014).

### 3.3 Study 1: Dictator and Ultimatum Games

#### *Design*

Based on prior work on moral standing for sentence structure (Khamitov et al., 2016) and MPT items for content (H. M. Gray et al., 2007), our manipulation was presented before participants partook in DG and UG as four different descriptions, that highlight low vs. high patency or agency traits.<sup>5</sup> The study was thus a 2 (Low vs. High) by 2 (Agency vs. Patency) between-participants factorial design.

#### **Participants and procedure**

We recruited participants on Amazon Mechanical Turk. Of the 202 participants, 131 were men (64.85%), 70 were women, and one person was of undisclosed gender. To report the most prominent age, race, and educational level categories, 101 (50%) were between 25 and 34 years of age, 154 identified as White (76.24%), and 135 had some college education or above (66.83%). The survey call stated that participants will partake in a task of distributing 20 tickets between themselves and a machine agent. Tickets entered them into a lottery for an additional \$10. Through a survey link, participants first read the informed consent form, answered demographic and emotion questions, and were randomly assigned one of the four conditions, with accompanying attention check questions that followed the description of a machine.

We called DG round one and UG round two, in order to not refer to these games by their known names. Participants had to read instructions about DG, which stated that they have "a higher chance of winning the lottery with more tickets." This was followed by attention check questions, before participants allocated tickets to the agent in DG. Then participants were asked about their emotion states. After that, instructions about round two (UG) followed that stated that the machine "can accept or reject your offer [...] (and that the machine's) rejection leads to zero tickets for both of you." After the attention

<sup>5</sup> Condition 1: It neither feels emotions nor reacts to the emotions expressed by others. Neither can it reason about how its actions and emotional expressions impact other people's emotions. Condition 2: It neither feels emotions nor reacts to the emotions expressed by others though it can reason about how its actions and emotional expressions impact other people's emotions. Condition 3: It feels emotions and reacts to emotions expressed by others, but it cannot reason about how its actions and emotional expressions impact other people's emotions. Condition 4: It feels emotions and reacts to the emotions expressed by others. It can also reason about how its actions and emotional expressions impact other people's emotions. In sum, (1) the machine *does not have a complex disposition to think, feel, and reflect* (low-agency, low-patency) vs. (2) *has a complex disposition to think and reflect, but cannot feel* (high-agency, low-patency) vs. (3) *has a complex disposition to feel, but cannot think or reflect* (low-agency, high-patency) vs. (4) *has a complex disposition to think, feel, and reflect* (high-agency, high-patency).

check questions, participants were asked to allocate tickets to the machine, given the new information that the machine can now overturn offers to the loss of both players. After DG and UG, the following measures were taken: MPT (H. M. Gray et al., 2007), stereotype content model questions (Fiske, Cuddy, Glick, & Xu, 2002)<sup>6</sup> the moral standing scale (Khamitov et al., 2016; Piazza et al., 2014), emotion states (de Melo & Gratch, 2015; Haidt, 2003; Skoe, Eisenberg, & Cumberland, 2002), the moral identity questionnaire (Black & Reynolds, 2016), and the inclusion of the other in the self (IOS) scale (Aron et al., 1992). We only report relevant measures in our results, in that the result of the moral identity questionnaire was not included. All participants received \$1.80 and one randomly chosen participant was awarded the extra compensation of \$10 at the end of the experiment.

### Manipulation check

Concerning perceived agency (MPT scale results), there was both a significant main effect of described agency of our text-based manipulation ( $F(1, 198) = 26.54, p < .001, \eta_p^2 = .118$ ) and a significant main effect of described patency ( $F(1, 198) = 14.92, p < .001, \eta_p^2 = .07$ ), whereas the interaction between agency and patency did not reach significance ( $F(1, 198) = 0.75, p = .39, \eta_p^2 = .00$ ). Participants perceived lower agency for the agent that could purportedly not reason ( $M = 2.88, SE = 0.17$ ) than when the agent was described as being able to reason ( $M = 4.09, SE = .17$ ). However, participants also rated the agent as lower in agency when it could not feel ( $M = 3.03, SE = 0.17$ ) than when the agent was described as being able to feel ( $M = 3.94, SE = .17$ ). Likewise, regarding perceived patency, there was a significant main effect of agency ( $F(1, 198) = 5.52, p = .02, \eta_p^2 = .03$ ) as well as a significant main effect of patency ( $F(1, 198) = 25.66, p < .001, \eta_p^2 = .12$ ), and the interaction between agency and patency was not significant ( $F(1, 198) = 0.59, p = .45, \eta_p^2 = .00$ ). Participants perceived lower patency for the agent that could purportedly not feel ( $M = 2.15, SE = 0.17$ ) than when the agent was described as being able to feel ( $M = 3.35, SE = .17$ ). However, participants also rated the agent as lower in patency when it could not reason ( $M = 2.47, SE = 0.17$ ) than when the agent was described as being able to reason ( $M = 3.03, SE = .17$ ). Given that agency and patency were highly correlated in the original MPT study that was conducted by Gray et al. (reported as “ $r(11) = .90, p < .001$ ” (H. M. Gray et al., 2007; Piazza et al., 2014)), we used the descriptions as intended.

<sup>6</sup> MPT dimensions conceptually relate to the stereotype content model (SCM). SCM deals with interpersonal perceptions of social group members based on two dimensions of *competence*, e.g., intelligent, competitive, confident, and *warmth*, e.g., friendly, good-natured, sincere (Fiske et al., 2002). Competence items evoke agency and warmth items are reminiscent of patency, though the aims of two scales differ (Haslam, 2012). SCM was not relevant for the current paper, but the trend after analysis was generally the same as MPT scales.

## Results

### Main analysis

For DG allocations, there was no main effect of agency ( $F(1, 198) = 0.21, p = .65$ ); however, there was both a near significant main effect of patency ( $F(1, 198) = 3.53, p = .062, \eta_p^2 = .02$ ) and a significant interaction between agency and patency ( $F(1, 198) = 6.26, p = .013, \eta_p^2 = .03$ ) for DG results. Whereas across patency conditions, participants gave less to the machine when it purportedly could not feel ( $M = 5.29, SE = 0.64$ ) than when the machine was described as being able to feel ( $M = 6.98, SE = .63$ ), this effect was driven entirely by the low agency condition ( $M = 3.96, SE = .90$  vs.  $M = 7.9, SE = .90$ ) and was absent in the high agency condition ( $M = 6.62, SE = .90$  vs.  $M = 6.06, SE = .90$ ).

In UG, there was both a significant main effect of agency ( $F(1, 198) = 3.90, p = .05, \eta_p^2 = .02$ ) and a significant main effect of patency ( $F(1, 198) = 7.58, p = .007, \eta_p^2 = .04$ ) on allocations, whereas the interaction between agency and patency did not reach significance ( $F(1, 198) = 2.12, p = .15, \eta_p^2 = .01$ ). Participants gave less to the machine when it could purportedly not reason ( $M = 8.63, SE = 0.51$ ) than when the machine was described as being able to reason ( $M = 10.04, SE = .51$ ). Likewise, participants gave less to the machine when it could not feel ( $M = 8.35, SE = 0.51$ ) than when the agent was described as being able to feel ( $M = 10.32, SE = .50$ ). Although covariance between allocations in UG and DG was high ( $F(1, 197) = 62.75, p < .001, \eta_p^2 = .24$ ), when controlling for DG outcome<sup>7</sup>, we observed the same pattern in UG; there was still both a significant main effect of agency ( $F(1, 197) = 4.02, p = .046, \eta_p^2 = .02$ ) and a significant main effect of patency ( $F(1, 197) = 4.31, p = .039, \eta_p^2 = .02$ ), and the interaction between agency and patency was not significant ( $F(1, 197) = 0.07, p = .80, \eta_p^2 = .00$ ). Thus, participants still gave less to the machine that could purportedly not reason ( $M = 8.71, SE = 0.44$ ) than when the agent was described as being able to reason ( $M = 9.97, SE = .44$ ). Likewise, participants allocated less to the machine that could not feel ( $M = 8.68, SE = 0.45$ ) than when the machine was described as being able to feel ( $M = 10, SE = .44$ ).

<sup>7</sup> We put in how much people gave the agent in DG as a covariate to control for its affect in analyzing UG outcomes

### Exploratory analysis

Our ANOVA analysis showed that people highly related to the agent (IOS) based on its manipulated patency, i.e., how much emotional behavior the agent showed ( $F(1, 198) = 6.99, p = .009, \eta_p^2 = .03$ ). But, agency and the interaction between agency and patency were not significant ( $F_s < .85, p_s > .36$ ). An agent described to have feelings was more relatable ( $M = 3.01, SE = .19$ ) than an agent that could not have emotions ( $M = 2.3, SE = .19$ ). As for the agent's moral standing, significance was found in regard to its manipulated agency ( $F(1, 198) =$

6.60,  $p = .011$ ,  $\eta_p^2 = .03$ ) and patience ( $F(1, 198) = 5.17$ ,  $p = .024$ ,  $\eta_p^2 = .03$ ). Their interaction neared significance ( $F(1, 198) = 3.65$ ,  $p = .06$ ,  $\eta_p^2 = .02$ ). The agent was granted higher moral standing when it could feel ( $M = 4.20$ ,  $SE = .17$ ) compared to when it could not feel ( $M = 3.65$ ,  $SE = .17$ ). Also, its high cognitive capacity contributed to greater moral standing ( $M = 4.23$ ,  $SE = .17$ ) compared to when the agent had low cognitive capacity ( $M = 3.62$ ,  $SE = .17$ ).

We lastly explored people's emotion states via repeated measure analyses of covariance (covariates being UG and DG outcomes) with agency and patience as between subject conditions. Changes in reported disgust and compassion over time was significantly linked to patience. There was a significant effect of patience on disgust over two rounds of DG and UG ( $F(1,196) = 6.25$ ,  $p = 0.01$ ,  $\eta_p^2 = .03$ ). If people got an unemotional machine, disgust went down ( $M = 1.98$ ,  $SE = .16$  to  $M = 1.77$ ,  $SE = .16$ ). If they interacted with an emotional machine, disgust went up (from  $M = 1.99$ ,  $SE = .16$  to  $M = 2.17$ ,  $SE = .16$ ). Compassion over two rounds also was affected by patience ( $F(1,196) = 4.32$ ,  $p = 0.04$ ,  $\eta_p^2 = .02$ ). If participants interacted with an unemotional machine, compassion went up across DG and UG rounds ( $M = 2.40$ ,  $SE = .18$  to  $M = 2.56$ ,  $SE = .18$ ). On the other hand, if participants were assigned to an emotional machine, compassion went down ( $M = 3.00$ ,  $SE = .18$  to  $M = 2.76$ ,  $SE = .19$ ).

Anger showed significance for agency, accompanied by a significant agency and patience interaction. Reported anger was significant for agency ( $F(1, 196) = 7.55$ ,  $p = 0.007$ ,  $\eta_p^2 = .04$ ) and for agency and patience interaction ( $F(1, 196) = 5.04$ ,  $p = 0.03$ ,  $\eta_p^2 = .03$ ). If the machine was reported to have low agency, anger went down over two rounds ( $M = 1.95$ ,  $SE = .14$  to  $M = 1.91$ ,  $SE = .16$ ), but anger went up if the machine was reported to have high agency ( $M = 1.76$ ,  $SE = .14$  to  $M = 2.1$ ,  $SE = .16$ ). As for the interaction, in the low agency condition, reported anger went up when coupled with low patience ( $M = 1.75$ ,  $SE = .21$  to  $M = 1.88$ ,  $SE = .24$ ); in contrast, anger levels *decreased* towards the low agency and high patience machine ( $M = 2.15$ ,  $SE = .21$ , to  $M = 1.93$ ,  $SE = .23$ ). In the high agency condition, anger levels went up for both low patience ( $M = 1.78$ ,  $SE = .20$  to  $M = 1.98$ ,  $SE = .23$ ) and high patience ( $M = 1.75$ ,  $SE = .20$ , to  $M = 2.22$ ,  $SE = .23$ ) conditions.

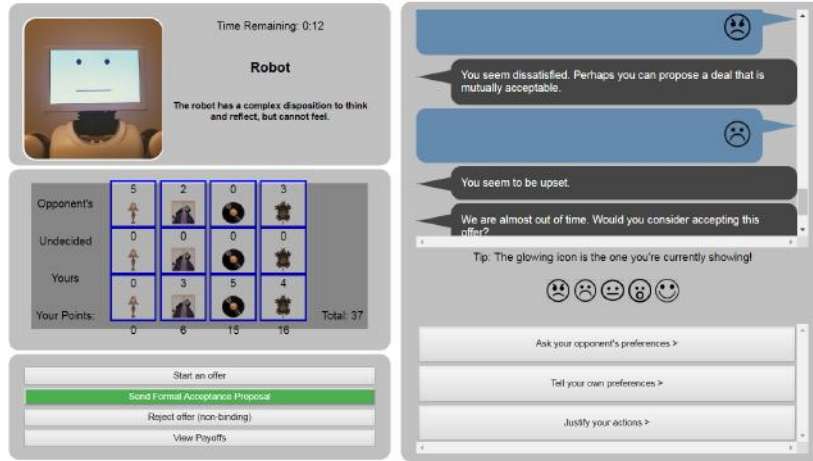


Figure 3.2: Negotiation interface in Study 2

Robot type	Description	Dialog
Low-Agency Low-Patency	<i>The robot does not have a complex disposition to think, feel, and reflect.</i>	<b>"Preparing offer." "Affirmative." "Does not compute."</b>
Low-Agency High-Patency	<i>The robot has a complex disposition to feel, but cannot think or reflect.</i>	<b>"I like this!" "Yay! I'm happy." "Oh...I'm sad..."</b>
High-Agency Low-Patency	<i>The robot has a complex disposition to think and reflect, but cannot feel.</i>	<b>"This is the most logical offer." "I inferred that you would accept this deal." "You seem to be upset."</b>
High-Agency High-Patency	<i>The robot has a complex disposition to think, feel, and reflect.</i>	<b>"I'm going to make this offer." "I feel so good about negotiating with you!" "Oh...Your sadness makes me feel sad..."</b>

Table 3.1: Agent types and excerpts from their descriptions and dialogues in Study 2.

### 3.4 Study 2: Negotiation

#### Design

Our agent was a virtual robot that was simple in appearance (Fig. 3.2), without any gender, race, or other highly anthropomorphic traits that may trigger people's biases (Bailenson et al., 2003; Dotsch & Wigboldus, 2008; Siegel et al., 2009), which helped to drive the perception of its mind based on its behavior rather than its looks. We used a configurable negotiation platform called IAGO for designing custom negotiation experiments. It features emotional communication (participants can click on different emojis to send to an agent; see Fig. 3.2), as well as customizable agents (e.g., agents' pictures can have different emotional expressions as reactions to people's behavior) (Mell & Gratch, 2017).

We again employed a between-participants factorial design of 2 (Low vs. High) by 2 (Agency vs. Patency) dimensions. Agency and pa-



tiency were manipulated in two ways. There were descriptions of the agent presented before the negotiation and shortened versions of descriptions appeared next to the picture of the agent (Fig. 3.2) during the experiment. These descriptions were the same as Study 1. In addition, we designed dialogues, i.e., how it “talked” (Table 3.1 lists excerpts). We used the items of the MPT scale (H. M. Gray et al., 2007) to construct the dialogues, as we did with descriptions. To illustrate, one agency item, “the robot appears to be capable of understanding how others are feeling” was translated to the agent having an awareness of the participant’s emotion states during the negotiation, e.g., a “sad” emoji from the participant resulted in “you seem to be upset” message from the high-agency low-patency agent while the agent’s expression remained neutral (Fig. 3.2). This suggests high-agency, but does not directly translate to a complete lack of emotional capacity (the agent is aware of the other player’s emotion states), even though the description stated it “cannot feel”.

We attempted to imbue the high-agency low-patency agent with an awareness of others’ emotions, e.g. “you seem to be upset”, whilst not being emotionally expressive itself, which are two different, but often conflated, design elements of affective virtual agents. In contrast, the low-agency low-patency agent did not use emotional language or expressions (static neutral face) and always responded to participants’ emojis with the statement “does not compute”. Hence, unlike prior work (de Melo, Gratch, & Carnevale, 2014), our agency and patency manipulation separated an agent’s awareness of displayed emotions (agency) from actually feeling emotions (patency). We imbued agency and patency features into agents’ descriptions and dialogues that occur *over time* in a negotiation (Table 3.1), which is how we carefully manipulated the mind dimensions compared to prior research (de Melo, Gratch, & Carnevale, 2014).

As a reminder, only dialogues and descriptions differed between agents (Table 3.1); the negotiation tactic was the same for all agents, for we are interested in the effects of MPT dimensions. Items for all negotiations were also the same with 7 clocks, 5 crates of records, 5 paintings, and 5 lamps, with different values per item per player for records and lamps. All agents began the negotiation by proposing the same starting offer (Table 3.3). The negotiation structure was partially integrative and partially distributive, meaning that half of the items were equally valuable to both players (distributive) while the other half of items had different values for players (integrative). This allows players to potentially “grow the pie” in a cooperative fashion through in-game communication while still playing competitively. Before the negotiation, partici-

pants were informed only about what they preferred. They were told prior to the experiment that one person who earned the highest points against the agent would get \$10 as a bonus prize.

	Clocks	Records	Paintings	Lamps
Robot	4	1	2	3
Human	4	3	2	1

Table 3.2: Points per item

All agents' negotiation strategy was based on the *minimax* principle of minimizing the maximal potential loss (Mell & Gratch, 2017); agents adjusted their offers if participants communicated their preferences, and strove for fair offers, while rejecting unfair deals. Agents did not know participants' preferences, but assumed an *integrative* structure. At the start, an agent made a very lopsided first offer (as a form of "anchoring") as shown in Table 3.3: it took almost all clocks (equally the most valuable item for both players), it allocated more lamps to itself (more valuable for itself) and gave more records to the participant (more valuable for the participant), and equally distributed the paintings (equally valuable item). This suggests that negotiators can cooperate and compete, potentially to enlarge the pie for both.

	Clocks	Records	Paintings	Lamps	Pts.
Robot	6*4	0*1	2*2	4*3	40
Undecided	1	1	1	1	
Human	0*4	4*3	2*2	0*1	16

Table 3.3: Agents' starting offer in Study 2: In all conditions, agents made the same, lopsided first offer as displayed. There were undecided items, one of each type. Points per item differed, thus the calculation stands as item \* points = total points.

### Participants and procedure

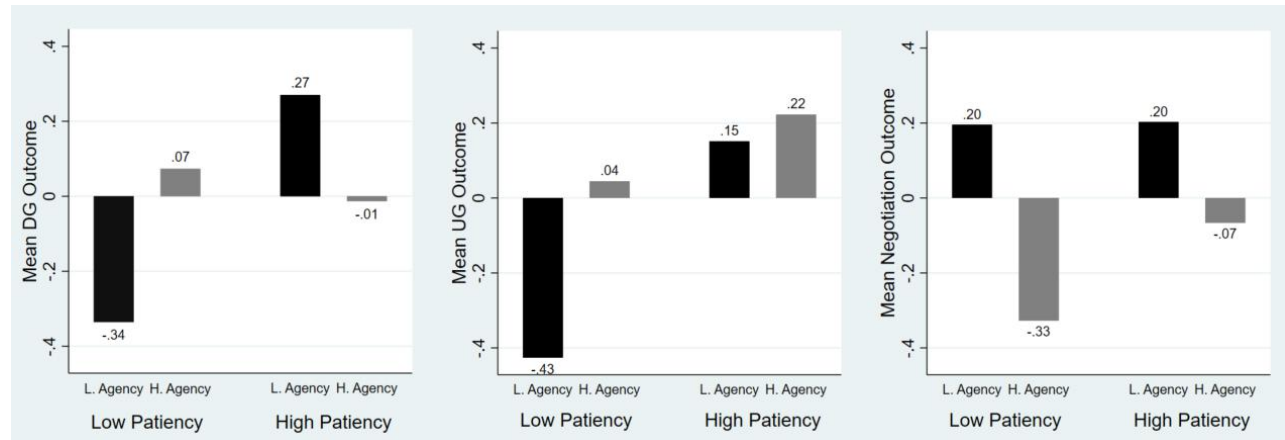
226 participants residing in the U.S. were recruited on Amazon Mechanical Turk. We had 135 men (59.7%), 90 women, and 1 of undisclosed gender. Participants were all over 18 years of age. 53.5% were between the ages of 25 and 34 (121 participants). Participants got a link to the survey that first contained the informed consent form, questions on participants' current emotion states and demographic information. Then participants read the description of an agent based on the randomly assigned condition (Table 3.1) and answered attention check questions about the description. After that, they read the instruction about the negotiation task, followed by additional attention check questions about the task, which they had to pass to go to the negotiation interface. They had up to 6 minutes to engage in a negotiation of four different goods (Table 3.2), and the count-down of time was displayed on the interface (Fig. 3.2). Upon completion of the negotiation, participants finished the second part of the survey of our measurements. We deployed the same measurements as the first study (Section 3.1.1). Further, we asked additional questions on whether or not participants made concessions to the agent and if the agent did

anything unexpected. We only report relevant measures in our results. Participants were compensated \$3 for their time, based on an estimate of 30 minutes to finish the entire survey and negotiation. One participant was randomly selected and awarded the \$10 bonus prize, after the experiment was completed.

### Manipulation check

Both of our experimental manipulations affected perceived agency; that is, there was both a significant main effect of agency ( $F(1, 222) = 35.68, p < .001, \eta_p^2 = .14$ ) and a significant main effect of patience ( $F(1, 222) = 53.42, p < .001, \eta_p^2 = .19$ ) on perceived agency, whereas the interaction between agency and patience did not approach significance ( $F(1, 222) = .60, p = .44, \eta_p^2 = .003$ ). Participants perceived lower agency for the agent that could purportedly not reason ( $M = 2.89, SE = .14$ ) than the agent described to have high ability to reason ( $M = 4.01, SE = .13$ ). But, people also rated the agent as lower in agency when it did not have affective capacity ( $M = 2.77, SE = .13$ ) than when the agent could have emotions ( $M = 4.14, SE = .13$ ). In contrast, only manipulated patience significantly affected perceived patience ( $F(1, 222) = 71.24, p < .001, \eta_p^2 = .24$ ); the effect of agency on perceived patience only approached significance ( $F(1, 222) = 2.57, p = .11, \eta_p^2 = .01$ ), and the interaction did not approach significance ( $F(1, 222) = .001, p = .99, \eta_p^2 = .00$ ). Participants rated the agent as lower in patience when it could not feel ( $M = 1.88, SE = .13$ ) than when the agent was described as being able to feel ( $M = 3.44, SE = .13$ ).

### Results



Due to inattention during the negotiation session, 78 participants were excluded as outliers for negotiation-related analyses. For user points,

Figure 3.3: Agents' standardized scores across DG, UG, and negotiation as outcomes, over low and high agency and patience.

there was a significant main effect of agency ( $F(1, 143) = 4.35, p = .04, \eta_p^2 = .03$ ); participants got more in the negotiation when the agent was described as being able to reason ( $M = 28.825, SE = .67$ ) than when the agent was described as not being able to reason ( $M = 26.69, SE = .77$ ). No other effects approached significance ( $F_s < .50, p_s > .48$ ). For agent points, there was also a significant main effect of agency ( $F(1, 143) = 6.68, p = .01, \eta_p^2 = .05$ ); agents got less in the negotiation when it was described as being able to reason ( $M = 34.06, SE = .76$ ) than when the agent was described as not being able to reason ( $M = 37.05, SE = .87$ ). No other effects approached significance ( $F_s < .23, p_s > .63$ ). Figure 3.3 displays the agent's end outcomes in DG, UG, and negotiation via standardized scores for comparisons; low-agency agents had the best outcomes in negotiations. Thus, the positive effect of agency on user points and the negative effect of agency on agent points cancelled out, such that the effect of agency on joint points was not significant ( $F(1, 143) = 1.66, p = .20$ ); no other effects approached significance ( $F_s < .58, p_s > .44$ ). Further, the effect of agency on the initial offer was not significant ( $F(1, 143) = .49, p = .49$ ); no other effects reached significance ( $F_s < 2.7, p_s > .10$ ).

Process measures capture *how* participants played against the agent and are important to negotiations. The effect of agency on game end time neared significance ( $F(1, 143) = 3.62, p = .059, \eta_p^2 = .03$ ); participants took longer if the agent was described as not being able to reason ( $M = 296.88, SE = 13.36$ ) than when the agent was described as being able to reason ( $M = 263.14, SE = 11.67$ ). But, this effect was driven entirely by the low-patency condition, as per a significant interaction ( $F(1, 143) = 5.38, p = .02, \eta_p^2 = .04$ ). The main effect of patency did not approach significance ( $F < .01, p > .99$ ). There was a parallel pattern for number of rejected offers. We saw a significant effect of agency on number of times users rejected offers ( $F(1, 143) = 9.50, p = .002, \eta_p^2 = .06$ ); participants were more likely to reject an offer if the agent was described as not being able to reason ( $M = .72, SE = .11$ ) than when the agent was described as being able to reason ( $M = .29, SE = .09$ ). However, this effect was again driven entirely by the low-patency condition, as per a significant interaction ( $F(1, 143) = 5.85, p = .02, \eta_p^2 = .04$ ). The main effect of patency did not reach significance ( $F < 2.32, p > .13$ ).

Participants chose to display the happy emoji significantly more when the agent was described as being able to feel ( $M = 1.25, SE = .18; F(1, 143) = 8.14, p = .005$ ) than when the agent was described as not being able to feel ( $M = .88, SE = .20$ ). No other effects reached significance ( $F_s < 1.92, p_s > .17$ ). Likewise, participants also chose to display the

surprise emoji significantly more when the agent was described as being able to feel ( $M = .47$ ,  $SE = .07$ ;  $F(1, 143) = 4.54$ ,  $p = .04$ ) than when the agent was described as not being able to feel ( $M = .25$ ,  $SE = .08$ ). No other effects reached significance ( $F_s < 1.60$ ,  $p_s > .21$ ). No other effects for any other emoji emotional display reached significance ( $F_s < 1.95$ ,  $p_s > .17$ ).

There were a few messages that participants sent to the agent (pre-set messages in the user interface) that were significantly used. Participants chose to convey the message “it is important that we are both happy with an agreement” more when the agent was described as being able to feel ( $M = .36$ ,  $SE = .06$ ;  $F(1, 143) = 5.18$ ,  $p = .02$ ,  $\eta_p^2 = .04$ ) than when the agent was described as not being able to feel ( $M = .16$ ,  $SE = .07$ ). No other effects approached significance ( $F_s < .03$ ,  $p_s > .85$ ).

The interaction between agency and patience significantly affected how often participants chose to convey the message: “I gave a little here; you give a little next time” ( $F(1, 143) = 4.25$ ,  $p = .04$ ,  $\eta_p^2 = .03$ ). People sent this message the most to the high patience, low agency agent ( $M = .158$ ,  $SE = .04$ ) and the least to the agent that was described to neither feel nor display cognitive thinking ( $M = -1.43$ ,  $SE = .05$ ). No other effects reached significance ( $F_s < 2.87$ ,  $p_s > .09$ ). There was also a significant interaction between agency and patience for this message “This is the last offer. Take it or leave it” ( $F(1, 143) = 3.88$ ,  $p = .05$ ,  $\eta_p^2 = .03$ ). The message was shared the most with the agent that was low in agency, but high in patience ( $M = .08$ ,  $SE = .03$ ) and the least with the agent that was high and both agency and patience ( $M = -3.5$ ,  $SE = .02$ ). No other effects reached significance ( $F_s < .85$ ,  $p_s > .36$ ). No other effects for any other message options reached significance ( $F_s < 2.17$ ,  $p_s > .14$ ).

### Exploratory analysis

Only manipulated patience significantly affected psychological distance (IOS) from the agent ( $F(1, 222) = 29.1$ ,  $p = .002$ ,  $\eta_p^2 = .04$ ); the effect of agency on IOS and the interaction did not reach significance ( $F_s < 1.16$ ,  $p_s > .28$ ). Participants reported that they identified with the agent more when the agent was described as being able to feel ( $M = 2.86$ ,  $SE = .16$ ) and that the agent was more distant from them psychologically when it could not feel ( $M = 2.14$ ,  $SE = .16$ ). Only manipulated patience significantly affected moral standing ( $F(1, 222) = 17.81$ ,  $p < .00001$ ,  $\eta_p^2 = .07$ ); the effect of agency on moral standing and the interaction did not reach significance ( $F_s < 1.53$ ,  $p_s > .22$ ). Participants rated the agent as lower in moral standing when it could not feel ( $M =$

3.08,  $SE = .16$ ) than when the agent was described as being able to feel ( $M = 4.03$ ,  $SE = .16$ ).

### 3.5 General discussion

Our article concerns how the imbued mind of agents based on MPT dimensions influence the results of DG, UG, and negotiation as human-agent interactions. In Study 1, the agent's described patience affected the allocation scheme in DG, with an interaction between agency and patience; in UG, described agency and patience influenced allocations to the agent (as in (de Melo, Gratch, & Carnevale, 2014)). Yet unexpectedly in negotiations, we only noted a significant effect of agency, in a different direction than anticipated. Low-agency agents ended with higher scores (Fig. 3.3) and also had longer negotiation periods. In comparison, high-agency agents had lower scores, particularly if they also had low-patience ("Spock"-like high-agency, low-patience agent), while negotiations themselves were shorter. The results on negotiation outcomes and processes, two paradigmatic measures in negotiation research (L. L. Thompson et al., 2010), did not align with our hypotheses, while DG and UG results echoed prior research (de Melo, Gratch, & Carnevale, 2014).

Compared to DG and UG, the interactive nature of negotiations means they allow for people to adjust their perception of non-human agents' minds. There are three premises on how changing perceptions may happen. First, people have preconceived beliefs about virtual agents' minds; agents are seen to have low-order theory of minds (H. M. Gray et al., 2007; Waytz, Gray, et al., 2010) (at least presently) even if people interact with agents socially (Blascovich et al., 2002; Nass et al., 1994). Second, the perceived mind of an agent can be adjusted, be it through patience (affective richness (K. Gray & Wegner, 2012)) or agency (behavioral intentionality (Levin et al., 2013)). Third, negotiations require cognitively effortful participation that involves theory of mind reasoning (de Weerd et al., 2017; Gratch et al., 2015), especially when it comes to mixed-motive negotiations (de Weerd et al., 2017; Pruitt & Kimmel, 1977). Through negotiations, an agent's behavioral intentionality can be called into question, providing people opportunities to reformulate an agent's degree of conferred mind.

All agents adjusted their offers in the same way if participants communicated about preferences (Mell & Gratch, 2017), so they appeared to calculatively negotiate though we did not implement any sophisti-

cated AI. Yet surprisingly, our low-agency agent did well against participants, interactively over time, though the high-agency agent did poorly against human participants that do have a higher degree of mind. Participants' behavior suggests that the common belief that technological agents have low-agency and low-patency (H. M. Gray et al., 2007; Waytz, Gray, et al., 2010) was called into question for low-agency agents; *perceived patency was insignificant to negotiation results.*

The disconnected nature between our low-agency agent's dialogues and descriptions vs. its negotiation style (mixed-motive games often require higher-order theory of mind) potentially called into question what the agent was "up to". People were potentially investigating the agent's behavior rather than focusing on the negotiation. When people cannot easily guess what an agent desires or intends to do, i.e., predict its intentional stance (D. Dennett, 1989), people may analyze why the agent was behaving in an unpredictable way. Participants with low-agency agents thus may have applied their higher theory of mind reasoning for investigating the agent itself, e.g., why it talks like an ATM or a child while playing a sophisticated game, while guessing at what such an agent would want from a negotiation. Thus, participants' priority would then be less about game strategies, but on investigating and questioning their bias—the inability of technology to have a human-like mind.

An agent that was described to be less cognitively intelligent (low-agency) interacted with participants in a cognitively taxing task (negotiation over goods), and this disjuncture gave people reasons to doubt their beliefs over time, i.e., we manipulated an agent's behavioral intentionality (Levin et al., 2013). Participants' assumed "winning" strategy could have drifted from point-based calculations as the time passed or it was initially assumed to not be just about item points. For one, emotional capacity of agents in Study 2 affected the outcome in an unexpected manner. Though people utilized more emotive messages and emojis with high-patency agents, this behavior did not influence outcomes since perceived patency did not impact the game. Potentially there was more "noise" to interpret when people interacted with high-patency agents—not only do they have to figure out game mechanics in terms of item values, but participants may have assumed that the agents' emotional capacity was for strategic reasons, even though agents' offer strategies were not affected by emotional communications from players. Emotions matter in how people take part in negotiations (Barry et al., 2004; Van Kleef, De Dreu, & Manstead, 2004), so people may have assumed that agents' emotions also served some purpose.

Qualities such as an agent’s emotions, moral standing and relatability are in essence, distracting points when it comes to game mechanics. Yet, these distractors could have (wrongly) gained greater traction as the negotiation continued over time, especially since harm salience regarding a moral patient increases with time pressure (K. Gray et al., 2014). Thus, by perceiving other minds over time, people can become sensitive to not only their own suffering as moral patients (Sanfey et al., 2003; Van’t Wout, Kahn, Sanfey, & Aleman, 2006), but also to the suffering of others, even when they are machine opponents (Bloom, 2017; K. Gray et al., 2014; Hewig et al., 2008). We find that identifying with a technological moral patient via manipulated mind can change people’s behavior towards it. Our exploratory analyses on IOS and moral standing contribute to this interpretation, and their relation to game outcomes are summarized in Table 3.4.

	Outcome	Moral standing	Relatability
DG	P + I	A + P + I	P
UG	A + P	A + P + I	P
Nego.	A	P	P

Table 3.4: The impact of manipulated agency and patency on outcomes, moral standing, and relatability (IOS) in Studies 1 and 2. Agency is denoted as A, patency as P, and their interaction as I.

We relate to others by seeing ourselves in them (Aron et al., 1992). And we utilize our own minds (Krämer et al., 2012) to relate to our own and others’ affective and cognitive capacities. These are summated as two dimensions of the mind, i.e., agency and patency (H. M. Gray et al., 2007), which builds on ample research that emotions and cognition mutually influence each other in driving behavior (Damasio, 2006; J. Greene & Haidt, 2002; J. D. Greene, Nystrom, Engell, Darley, & Cohen, 2004). However, only patency, the perceived propensity to feel emotions, significantly contributed to how much people identified with agents across DG, UG, and negotiation (Table 4). Hence, studies 1 and 2 have the same trend regarding relatability (IOS) (Aron et al., 1992). People related to agents’ patency, even if agency may drive their strategic decision-making when greater “mind reading” is required as economic exchanges become more complex.

Interestingly, agent’s moral standing followed a different pattern from IOS: in DG and UG, both described agency and patency affected moral standing, but only imbued patency impacted the agent’s moral standing in negotiations. The outcomes of strategic games increasingly hinge on perceived agency for as the exchange becomes more complex (from DG and UG to negotiation). But no matter how complex an exchange is, people relate to an agent that has affective capacities. An agent’s moral standing may become more dependent on its perceived ability to feel when strategic exchanges become more complex. The agent may be seen more as a moral patient over time (K. Gray et



al., 2014), as the strategic exchange increasingly make people exercise greater human agency against the machine opponent that is then seen more as a moral patient (Fig. 3.1).

A novel implication is that mind perception may require theoretical revisions to account for *interactive* opinion formation about an agent's mind; negotiations provide a contextually different framework than a single instance evaluation of an agent's mind (as in DG or UG). Mind perception theory focused more on the latter case; it is about people's pre-existing beliefs at a single point in time and minds of various beings were judged through a survey (H. M. Gray et al., 2007). The novelty of our studies is that people seem to be revising their opinion of the agent's perceived mind over the course of a complex interaction. The human attribution of a mind in a machine may be misguided and can be revised; people can question their own beliefs through an interaction. Negotiations are potentially one of many interactive paradigms that can better clarify how people assess agents that display different degrees of having a mind in different ways over time. More relevantly, exploring other types of exchanges, e.g., purely integrative or distributive negotiations, can reveal in what ways an agent's perceived mind impact people as they attempt to understand whether or not a social agent is also an intentional agent.

### 3.6 Conclusion

We are far from having virtual agents that are truly intentional actors like humans. But, the degree to which agents are perceived to have agency and patiency, and what effect such manipulation has on us was observed in our studies. The DG outcome was influenced by perceived patiency of an agent, and the UG outcome was affected by perceived agency and patiency. Yet compared to single instance economic exchanges like DG or UG, interactive negotiations allowed us to catch a glimpse of how people react when they encounter agents that behave counter-intuitively, e.g., negotiating in an agentic manner without prescribed agentic traits. In negotiations, participants got more points against an agent with high-agency. In contrast, they did worse, took longer to play, and rejected more offers from a low-agency agent, as influenced by patiency. Patiency resulted in more emotional expressions from participants to the agent; people engaged more with emotional signals, i.e., emojis and messages.

As interactions require people to increasingly exercise more complex

theory of mind reasoning against non-human agents, e.g., from DG, UG, to negotiation, game outcomes depend more on agents' cognitive traits while their moral standing depends more on perceived affective traits (Table 4). Both agency and patiency contributed to an agent's moral standing after DG and UG, but people granted higher moral standing and related more to the agent through its patiency after negotiating with it. People's ability to relate to agents consistently is on whether they can have human-like feelings, regardless of people's own level of theory of mind required in an interaction.

Artificial emotions may uniquely contribute to machines' moral standing only when humans interactively act *against* machines with agency while concurrently, machines respond with traits of being moral patients, e.g., emotional reactions. We additionally conjecture that a virtual agent that sends unclear or mismatched signals that people have to interpret during a complex interaction like negotiation can lead people to reconsider agents' perceived minds, more so than in single-shot games like DG and UG that are not interactive. What we can conclude is that in attempting to comprehend a virtual agent's "mind", people react to its rational and emotional capacities in divergent ways, leading to noticeable differences in how they behave.



# 4

## *“You’re a robot, so you don’t feel so much”*

### 4.1 Introduction

If artificial agents are becoming a part of our societal fabric (Danaher, 2019) and are perceived to have minds of their own (K. Gray, Young, & Waytz, 2012; Waytz, Morewedge, et al., 2010), machines may be more frequently making moral decisions alongside us in the future. Hence, future machines are argued to be able to “perform deep moral and social reasoning about real-world problems” (Yang et al., 2018, p. 7). This is a process that is suggestive of a mind, but also a process in which a non-human agent can disagree with and attempt to influence a person’s moral choice— an experience we now encounter with other people, not agents. This may perhaps change in the future (Klincewicz, 2017; Wallach & Allen, 2008), when we accomplish the technical feat of building machines with minds. While this is a valuable research avenue, a more prescient project is on how we can be affected by machines with *perceived* minds and in what ways they can transparently communicate with us during moral interactions.

A longstanding issue is that what makes an interaction with technology *moral* or *immoral* is notably unclear, which leads to further ambiguity on whether a digital entity can ever be called “moral” compared to being, e.g., safe (van Wynsberghe & Robbins, 2019). Thus in our study in this chapter, we defined a moral interaction as when a human

and robot discuss an ethical dilemma, e.g., the trolley problem (Foot, 1967), with each other. In this, we considered the robot to be performatively equal to a human in having a moral status (Danaher, 2019) when it *conversationally demonstrates reflections on the moral elements of the dilemma and disagrees with a person's moral position with its own ethical position*.

We looked into how people *react* to a robot that *appears* to have a mind when expressing a moral opinion that is different from theirs, and if factors like visual reminders that one is talking to a mere robot affects this interaction. The motivation is that whether or not the fundamental aim of AI research, i.e., to create machines with minds (Haugeland, 1989) is possible, we need a better understanding on how we are affected by such machines. As the primary aim, we were concerned with people's judgement of a robot's moral status and people's moral decision-making process with a robot through a qualitative lens. The secondary exploration was on how to transparently remind people that the robot that shares its moral opinion is a mere machine with quantitative analyses. We pursued the following research questions. *How do people perceive a robot that attempts to influence their moral position? Do visual cues for transparency on a robot's mental states affect people's perception of the robot's mind and social attributes (as competence, warmth, and discomfort)?* We now present related literature, followed by our methods and results per study. Then a discussion and implications are presented.

## 4.2 Related work

The focus, as before, is on how people view an artificial agent and this affects them: I build on prior chapters on artificial emotions and the changing mind perception of an artificial agent during an interaction. I first provide a background literature on morality that broadly touches on emotions and moral status. Then I elaborate on transparency, in which I discuss explicability and trust of AI systems.

### *Morality and emotions*

How people perceive a non-human agent to have a mind, i.e., how it is perceived to *think* and *feel*, is critical in morally loaded situations. Perceived minds can, for instance, influence how people trust agents during interpersonal exchanges (de Melo, Gratch, & Carnevale, 2014). One way that people form the belief that an agent has a mind is via its emotional behavior (de Melo, Carnevale, et al., 2014). In render-

ing agents' artificial emotions, top-down (often labelled cognitive) and bottom-up (emotive<sup>1</sup>) designs are both helpful for morally relevant human-robot interaction (HRI) (Wallach & Allen, 2008).

<sup>1</sup> Emotions are bottom-up if we take them to be "feelings of patterned bodily changes" (Prinz, 2004, 2006, p. 34).

In currently deployed technical systems with morally relevant roles, e.g., care robots or weaponized drones, we see instances of top-down, cognitive engineering of morality, i.e. logic-based action, rather than bottom-up, evolutionary training, i.e., learning from repeated experience (Wallach & Allen, 2008). A "bottom-up" approach can be beneficial for artificial emotional complexity, for example, to design robots that demonstrate empathy (A. Lim & Okuno, 2015). Though many challenges remain to practically achieve this, emotionally responsive robots can bring about engaging HRI (Picard, 1995, 2003), which impacts moral interactions.

Emotions are argued to be important for a moral agent, even when they are merely mimicked (Prinz, 2004, 2006). For example, psychopaths are "parasitic" on genuine moral emotions of other humans though they themselves do not necessarily feel moral emotions (Prinz, 2006). Here, what is meant by parasitic is that moral emotions of others are necessary for a common moral language even if psychopaths are merely subscribing to others' moral conventions; amoralists, such as psychopaths, or anthropologists of a different culture, have the ability to fathom moral realities of others without sharing their moral emotions (Prinz, 2006). What about robots that also do not readily feel moral emotions?

For robots to have moral status in humans' eyes, as in being attributed with moral agency and patiency (see Fig. 3.1, p. 48), their affective and cognitive capacities are equal components to their moral standing (K. Gray, Young, & Waytz, 2012). People perceive minds in others, even non-humans, again based on the two dimensions of agency that encompasses cognitive traits, e.g., being goal-oriented, having memory, and patiency as affective traits, e.g., propensity to feel joy or anger (H. M. Gray et al., 2007; K. Gray & Schein, 2012). Regarding social attributes, *competence* traits, e.g., knowledgeable, capable, are related to perceived agency and *warmth* attributes, e.g., social, happy, are related to patiency (Carpinella, Wyman, Perez, & Stroessner, 2017; Cuddy et al., 2009). To simplify, competence and agency are generally about cognitive faculties and warmth and patiency are more about emotions.

One scenario that activates our cognitive and emotive neural circuitry is the footbridge dilemma. The footbridge dilemma is a variation of the trolley problem as a thought experiment that was originally intro-

duced by Foot (1967). Foot contrasted two hypothetical and contrived moral dilemmas to probe readers' intuitions. In the trolley dilemma, a hypothetical out-of-control trolley is about to kill five workers on a track, and a person is asked if she would pull a switch to divert the trolley to a different track on which one person is working; one person's death then can save more lives (Foot, 1967). In the footbridge dilemma, the same out-of-control trolley is about to kill five workers on a track, but this time a person is asked if she would *push* a human bystander off of a bridge to stop the trolley. Foot was concerned with distinguishing between allowing harm and actively causing harm (Foot, 1967) (the doctrine of double effect<sup>2</sup>), and people react differently based on this distinction.

In fMRI experiments with human participants, the footbridge dilemma triggered an emotional response in the brain<sup>3</sup>, but the trolley dilemma activated rational control which can override initial emotional judgments in favor of a consequentialist<sup>4</sup> cognitive process to save more lives (Cushman, Young, & Greene, 2010; J. D. Greene, 2007; J. D. Greene et al., 2001). Imagining personally harming someone by pushing this person to death is taken to be emotionally more costly than impersonally pushing a switch, though the ends are the same (instrumental death of one person to save five lives).

Human reactions to a robot's immoral acts vary depending on whether they are third-party observers or second-party interactants. From people's third-person point of view, robots are not necessarily seen as emotional entities (H. M. Gray et al., 2007) and are expected to make decisions that favor the greater good even though one life has to be sacrificed (Malle et al., 2015). From a second-person standpoint, people react as if robots are accountable for their actions (Short, Hart, Vu, & Scassellati, 2010), which may involve emotions from humans, but also robots. Not only do humans respond emotionally, e.g., being upset at a robot's cheating behavior (Short et al., 2010), but robots that use emotional expressions themselves are treated differently than robots that do not utilize emotions by human interactants (de Melo, Gratch, & Carnevale, 2014; Lee, Lucas, et al., 2019). A robot's perceived reasoning and affective capacities, and the perception of its mind itself, could change over time during a second-person interaction (as in Chapter 3).

### *Moral status*

Whanganui River in New Zealand was granted personhood in 2017.<sup>5</sup> In Bangladesh, all rivers got the status as persons in 2019. This means that when a river is harmed, for instance through pollution, it can

<sup>2</sup> The doctrine of double effect: Allowing harm, e.g., instrumental death of one person, for the greater good, e.g., five lives saved, is permissible as a side effect in comparison to actively causing harm in order to achieve a good end.

<sup>3</sup> It is important to point out the intrinsically linked nature of brain regions. In that, when we say "emotional response in the brain", it does not mean that there is a hub *only* for emotional processing. Regions for working memory and emotional processing do relate to one another, but thinking about morally loaded, personal situations can trigger the emotional processing region of the brain more so than the region associated with memory processing (J. D. Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). An issue is the technical flaw of fMRI analysis softwares outputting false positive of up to 70% that was fixed in 2015, which would be before the cited research here was done (Eklund, Nichols, & Knutsson, 2016). Another issue is the lack of external validity of highly stylized thought experiments like the trolley dilemma that do not reflect real-life moral problems (Bauman, McGraw, Bartels, & Warren, 2014).

<sup>4</sup> As a thought experiment, the dilemma is said to contrast deontological (rule-based morality) and consequentialist (utility maximizing, ends-over-means morality) ethics.

<sup>5</sup> BBC:  
<http://www.bbc.com/travel/story/20200319-the-new-zealand-river-that-became-a-legal-person>

legally sue entities that are responsible for the damage. While the accountability on who represents rivers at the institutional level can be difficult to pinpoint<sup>6</sup>, the aim is environmental protection by the means of granting specific bodies of nature the same legal status as humans towards the end of changing our actions that impact rivers. Protecting our environment for many people is an important moral responsibility, whether or not nation-states take on institutionally defined legal responsibility. But what about the status of artificial beings that we create?

In 2017, the European Parliament released the Resolution on Civil Law Rules of Robotics which stated that “at least the most sophisticated autonomous robots could be established as having the status of electronic persons [...] where robots make autonomous decisions or otherwise interact with third parties independently”.<sup>7</sup> Though we are far from having machines that can independently make moral decisions on their own, we already see personhood being granted to Sophia<sup>8</sup>, a robot with very limited AI that gained Saudi citizenship (Nyholm, 2020). A self-driving car perhaps has more complex AI, but a car is far from being seen as a candidate for personhood.<sup>9</sup> A robot then can be made to look like a person (or not) with varying degrees of AI; what it looks like from the outside as well as how it is constructed from the inside can both contribute to calling something a robot with AI (Nyholm, 2020).

Robots’ legal status can invite or signal moral status in a normative sense; institutions or governments ascribe legal status, e.g., a driver’s license to a car or citizenship to a humanoid robot. Yet, people’s behavior towards artificial agents can suggest that they have moral status in a descriptive sense. While people may know logically that robots are mere things, people may still treat robots as if they have some form of moral status via perceived agentic or patient capacities. For example, a robot that cheats during a game may be called out as acting unfairly; even if people logically know that they are playing against artificial agents, they may react emotionally *as if* robots should know better (Short et al., 2010). Such reactions are harder to account for when designing robots. Though the European Parliament states that designers should make sure that “robots are identifiable as robots when interacting with humans”, they also write that users should “respect human frailty, both physical and psychological, and the emotional needs of humans”. The Civil Law Rules of Robotics of 2017 notes emotional and psychological care as an individual’s responsibility.

Somehow “human frailty” is not a concern when we talk about a

<sup>6</sup> NPR:  
<https://www.npr.org/2019/08/03/740604142/should-rivers-have-same-legal-rights-as-humans-a-growing-number-of-voices-say-ye>

<sup>7</sup> Civil Law Rules of Robotics by the European parliament, 16 February 2017:  
<https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051EN.html> (accessed on 23 March 2020).

<sup>8</sup> Hanson Robotics:  
<https://www.hansonrobotics.com/>

<sup>9</sup> The state of Nevada in 2012 did issue the first license for a self-driving car (Bühler, 2015; Nevada, Department of Motor Vehicles, 2012). Though this does not designate the car to have moral status, it signals that it has the legal status of a driver.



toaster that just toasts. But when a toaster starts to talk or cheat, it should still be enough that it calls itself a toaster. Then if I am emotionally and psychologically affected by my strange, talking toaster that cheats me out of a good toast, I should keep my frail human self in check. Imagine that it is not a toaster, but a robot that disagrees with people during emotionally-loaded moral decision-making. Or even further, a talking doll or anime character that invites attachment; people can take emotional solace and romantic interest in artificial beings and thus see them as having some moral status (McArthur & Twist, 2017; Nyholm & Frank, 2017). Our human frailty and ascription of moral status to artificial agents is still our own business.

Ethics of AI that is coupled with robotics then is concerned with whether the onus falls solely on the individual when it comes to one's emotional and psychological health. When one *chooses* artificial love, one exploits one's own human frailty. Openness to vulnerability is at the heart of human morality, i.e., what Nussbaum calls "fragility of goodness" (M. C. Nussbaum, 2001). But should we be openly vulnerable towards artificial agents to the extent we may be towards other humans? We will not resolve this question here, but we raise the point that people are acting emotionally vulnerable towards machines, whether this is their choice or not (if the difference can be clearly made). The ability to suffer, the aspect that patiency covers (H. M. Gray et al., 2007; K. Gray & Schein, 2012), relates to moral status, for both machines and human users, i.e., we can suffer due to machines, but we can also perceive suffering in machines. The distinction is on whether artificial suffering, or more broadly artificial emotions, *should* be enough in granting things moral status, when people may grant machines moral status based on artificial emotions.

According to Midgley, "what makes creatures our fellow beings, entitled to basic consideration, is surely not intellectual capacity but emotional fellowship" (Midgley, 1996). Even Bentham shared that "the question is not 'Can they reason?' or 'Can they talk?' but 'Can they suffer?'" (Bentham, 1996 [±1789-1843]). But, when exactly emotional fellowship becomes relevant for HRI should be clarified before we approach norms as prescriptions vs. behavior. Specified to human behavior, we saw in Chapter 3 that when people can (and are asked to) act more agentially against a machine, e.g., negotiations, its perceived patiency comes to the fore when people judge its moral standing (Table 3.4). Hence a revision to the question 'Can they suffer?' is 'Do I act as if they suffer when I exercise my agency to make machines suffer, even if I believe machines cannot suffer like us?'

### *Transparency of systems and robots*

Transparency is recognized as a core ethical principle for AI in society. Also called explicability, the principle asks any artificial system to be upfront about how it works and who can be held responsible for the way it works (Floridi & Cowls, 2019; High-Level Expert Group on Artificial Intelligence, 2019). Transparency is broadly defined as “a way for the human and the machine to be on the same page with regard to goals, processes, tasks, division of labor within tasks, and overall intent-based approach toward the interaction” (Lyons et al., 2017, p. 128). Especially when people have to collaborate with a robot, transparency on the robot’s purpose and function can foster trust (Muir, 1987). However, there is a research gap on how to best link the AI mechanism (“under-the-hood” workings) of a robot to understandable explanations for humans, when considering dynamically changing environments involving multi-modal cues (Anjomshoe, Najar, Calvaresi, & Främling, 2019).

Communicating a robot’s purpose as transparency cues comes in many forms. Different methods of transparent communication impact people’s level of trust. Robots that transparently reveal why they made certain decisions, e.g., via visualized information, are taken to be more trustworthy and understandable compared to systems that do not clearly explain their decisions (B. Y. Lim, Dey, & Avrahami, 2009). There are other ways to visually explain how a robot works. Some examples include a robot’s decision tree that people can zoom in on (Brooks, Shultz, Desai, Kovac, & Yanco, 2010), querying interface (Lomas et al., 2012), and text combined with GUIs for dynamic decision-making (Mercado et al., 2016). Auditory cues are also relevant. A robot that verbalized its reasoning after taking action was found to gain people’s trust (N. Wang, Pynadath, & Hill, 2016). Usually, the highest level of transparency has been interpreted as giving the most comprehensive, logical explanation by a system, which translated to a high level of trust in the system (Lyons et al., 2017). While information overload is a worry with multi-modality, a study found that trust increased according to the transparency level at no cost to cognitive workload (Mercado et al., 2016). Transparency cues can be effective in garnering people’s trust, but there are two issues.

One issue is in how users can best *calibrate* their level of trust in machines; when people cannot accurately predict how a machine functions, what it can competently do, and where their responsibilities in decision-making lie, people can over-trust or under-trust robots (Lyons et al., 2017; Muir, 1987). In relation, a robot’s level of transparency can

counter-intuitively work against the original intention of helping people make appropriate decisions, i.e., automation bias can occur when people over-rely on systems' decisions or suggestions (Skitka, Mosier, & Burdick, 1999). Transparent explanations can thus worsen automation bias; transparency cannot mitigate people's over-trust and it has actually been found to exacerbate automation bias instead (Schaffer, O'Donovan, Michaelis, Raglin, & Höllerer, 2019). Transparency adds to over-trust when people already tend to rely on robots for certain tasks due to automation bias.

The second issue is that when there is no automation bias, providing further transparency can lower perceived accuracy of the robot, if it provides highly accurate information to begin with; additional information that people did not ask for may raise unnecessary concerns (Springer & Whittaker, 2018). But when a robot gives inaccurate information, transparency does add value in "damage control", i.e., explaining where it went wrong (Springer & Whittaker, 2018). Transparency then serves as a way to re-set people's initial expectations, for better or worse. Trust in robots erodes when people's expectations are unmet or violated, but transparency on why expectation violation occurred can uphold users' trust (Kizilcec, 2016).

To reiterate, transparency can solidify automation bias resulting in over-trust (Schaffer et al., 2019), yet over-explaining via transparency can make users question the accuracy of a system (Springer & Whittaker, 2018). In case mistakes are made, however, transparency cues help to explain why errors happened, which can preserve trust (Springer & Whittaker, 2018). Hence, "designing for trust requires balanced interface transparency—not too little and not too much" (Kizilcec, 2016, p. 4). Situational transparency may then be apt; the interface should adapt to the context and/or user. The challenge is in distinguishing which combinations of modalities and communication channels are well suited for people to comprehend the robot's mental states in specific situations; the robot's intent for transparent communication may not be correctly deciphered by users, especially in complex social settings (Anjomshoae et al., 2019).

When robots act as social agents, they have built-in affordances that allow them to adapt to us. Or, we as social agents easily adapt to artificial beings due to our built-in nature to socially react to anthropomorphized robots (Lottridge, Chignell, & Jovicic, 2011). A robot that acts socially, e.g., saying "hello", is treated in a human-like way, e.g., people say "hello" back to it (Moon & Nass, 1996; Nass et al., 1994). People adapt their behavior to social agents, e.g., they tend to help a

computer that was previously helpful (Fogg & Nass, 1997). In line with this, we can attribute social traits to a robot based on its appearance (Duffy, 2003; Fink, 2012; Walters, Syrdal, Dautenhahn, Te Boekhorst, & Koay, 2008), or its behavior (Cuijpers & Knops, 2015; Ruijten & Cuijpers, 2017; Salem, Eyssel, Rohlfing, Kopp, & Joublin, 2011). When a robot is equipped with eyes or other facial features, it can show emotional expressions. When it has a voice, it can have a spoken interaction. And when it has arms and legs, it can use them for gesturing. Even a robot that shows idle motions, e.g., swaying back and forth, is perceived as more social than when it does not show these motions (Cuijpers & Knops, 2015).

Behaviors that arise from a robot's gaze, voice, gestures, and/or facial expressions can be meaningful social cues that suggest human-likeness and mind, which are not only important for natural interactions (Duffy, 2003; Eyssel, Kuchenbrandt, & Bobinger, 2011; Lemaignan, Fink, Mondada, & Dillenbourg, 2015; Ruijten, Bouten, Rouschop, Ham, & Midden, 2014), but also for transparent communication. A robot's gaze facilitates cooperation as it signals attentiveness (Admoni, Datsikas, & Scassellati, 2014; Admoni & Scassellati, 2017). Further, non-embodied gaze, i.e., from smart speakers, do not induce social behaviors like turn-taking, unlike gaze from an embodied agent, i.e., a human-like robot, that people react socially to by practicing turn-taking (Kontogiorgos, Skantze, Abelho Pereira, & Gustafson, 2019). Moreover, when people are asked to follow a robot's gaze towards specific objects, they attribute greater intentionality to the robot compared to when it merely identified an object (Levin et al., 2013). This is perhaps why a robot museum guide that gazed at art to direct people's focus to it helped people remember more details about art (Karreman, Ludden, & Evers, 2019); it used gaze in a natural way for humans, which translated to intuitive trust for museumgoers. But when the robot used a visual cue (an arrow on a screen) to point at art rather than gazing at it, the robot was more positively viewed and people paid more attention to the robot itself (Karreman et al., 2019). Thus, effectively combining human-like social cues with machine-like transparency cues, e.g., GUI, is a challenge; humans easily attribute mind-related characteristics like intentions, beliefs, or desires to a robot's behaviors even when it was not designed to convey such attributes, and people's attribution of intention influence how transparency cues are understood (Anjomshoae et al., 2019; Hellström & Bensch, 2018; Karreman et al., 2019).

As aforementioned, showing *why* a robot took a certain decision is at the heart of *how* transparency can be displayed. When a robot is

transparent about its reasoning process, e.g, its series of mental states, *in situ*, people can perceive it to be more human-like and trustworthy than robots that do not offer such explanations (de Visser et al., 2012; Jian, Bisantz, & Drury, 2000) and such robots are more likely to have perceived minds (De Graaf & Malle, 2017; Hellström & Bensch, 2018; Premack & Woodruff, 1978). But beyond transparency via revealing a robot's cognitive reasoning, the under-realized path is transparency through emotional communication that social agents can provide. A nameless, faceless GUI interface (in a traditional sense) does not evoke social reactions in humans as much as a robot that can use its voice, facial expressions, and/or bodily movements to communicate.

People's trust in robots is affected by their looks and physical presence (Bainbridge, Hart, Kim, & Scassellati, 2008) and also robots' voice (Torre & White, 2021). In such ways, social robots can express themselves multi-modally to convey emotions that other uni-modal interfaces cannot easily do. Why does emotional communication matter for transparency? Schaffer and colleagues suggest that when people are over-confident or have deep-seated automation bias, "non-rational methods, such as appealing to emotion, may be the only avenue to accommodate the overconfidence" (2019, p. 248). A robot's GUI(s), voice, gaze, gestures, and/or facial expressions that are presented together can convey social and emotional communication, but more importantly, they can affect human interactants' emotions and decisions. Robots' artificial emotions can impact transparency, e.g., by focusing users' attention toward or away from relevant information, and hence multi-modal expressivity can be pertinent to moral interactions.

### *Research aims*

While designing AI with artificial morality is arguably indispensable in the responsible creation of intelligent machines (Allen, Smit, & Wallach, 2005; Danielson, 2009; Wallach & Allen, 2008), how we may be impacted by machines that behaviorally claim equal moral status, by for instance disagreeing with us on moral grounds, is not well understood. Our research hence does not focus on whether or not robots can or should be "moral" or what that could mean, conceptually or technically. Instead, our interest lies in factors that influence people's perception of a robot that discusses a moral dilemma. We explored the following question qualitatively: how do people perceive a robot that attempts to influence their moral position? We supplemented this with quantitative measures in asking the following. Does visually displaying robot's step-by-step mental states on a screen during the moral discussion affect people's judgment of the robot's perceived mind and

social attributes (as competence, warmth, and discomfort)? We turn to methods and results for our study.

### 4.3 Methods

Prior studies (Chapter 3) indicated that the robot’s emotional displays (like the usage of emotional language) does affect its perceived mind. Here, we did not manipulate the robot’s behavior to be emotional in order to prioritize the emotional situation and debate shared by participants and the robot. The situation was an *emotionally loaded moral scenario*, i.e., the footbridge dilemma on whether or not one would push vs. not push a person rather than flipping a switch as in the trolley dilemma. In this, a robot *disagreed* with all participants on whether or not one should push. We explored the following question qualitatively: how do people perceive a robot that attempts to influence their moral position? We supplemented this with quantitative measures in asking the following. Does visually displaying a robot’s step-by-step mental states on a screen during the moral discussion affect people’s judgment of the robot’s perceived mind and social attributes (as competence, warmth, and discomfort)?

Our main manipulation was on whether or not there was a screen next to the robot as our independent variable for our quantitative analysis. The screen displayed the robot’s transition through its “mental states”, e.g., replying or thinking. These visual transparency cue may be especially important during moral HRI when a robot attempts to influence people’s moral decisions. The dependent variables were thus people’s perception of robot debater’s perceived mind, social attributes, and trustworthiness for between-subject comparisons. We used the simplified mind perception scale<sup>10</sup> (Ruijten et al., 2014; Waytz, Morewedge, et al., 2010) and explored other factors, specifically the robot’s perceived social attributes (Carpinella et al., 2017) and people’s trust in the robot (Jian et al., 2000).

<sup>10</sup> All measurement instruments are found in Appendix A.

Qualitatively, we were interested how participants experienced the scenario and the debate. We looked into people’s responses to the robot. Participants’ in situ behavior with the robot was observed via recorded videos and during the experiment from another room. Our analysis of participants’ behavior was corroborated by their written responses to open-ended questions on their opinion about the interaction post hoc. In this, we prioritized participants’ first-person views and behavior whilst taking an interpretive stance as researchers (Van Manen, 2016).

Participants' behavior in their interaction with the robot includes our interpretations on their use of body posture, gestures, gaze, and verbal reactions to the robot; their first-person thoughts regarding the experiment as written responses further adds to the richness of our qualitative data (Van Manen, 2016). People's corporeal, situational, and reflective experiences frame our understanding of quantitative analyses conducted. Thus, our qualitative and quantitative analyses supplement each other as concurrent triangulation (J. Creswell, Clark, Gutmann, & Hanson, 2008; J. W. Creswell & Creswell, 2003; Patton, 1999). As a strategy, concurrent triangulation means that we collected and analyzed our quantitative and qualitative data in parallel; each data type is interpreted in relation to other data.

In integrating and comprehending our qualitative data, we performed thematic analysis, which is flexible for various research agendas and can be used in conjunction with many schools of qualitative research (Braun & Clarke, 2006). We hence inductively looked for patterns to start. Our aim was to decipher "meaning units", i.e., units of analysis that capture an underlying phenomenon on the basis of participants' body language, tone of voice, and/or text, that are both implicit and explicit (Giorgi, 2012). In this process, both qualitative and quantitative analyses contributed to our understanding, in that the researchers openly engaged with all collected material simultaneously to best corroborate meaning units. Then, our meaning units were grouped and rearranged into relevant sub-themes and themes, which are elaborated on in the results section.

### *Participants and design*

Seventy-one participants (31 males and 38 females, 2 did not indicate their gender,  $M_{\text{age}} = 22.0$ ,  $SD_{\text{age}} = 3.1$ ) were recruited from Eindhoven University of Technology's participant database, after the study design passed an ethical review. 59 out of 71 participants had previous experience with programming and 55 had interacted with a robot before. Our between-subjects factor consisted of transparency vs. no transparency conditions. In the transparency condition ( $N = 35$ ), extra information about the robot's current mental state was presented to participants on a computer screen; see Figure 4.1. The robot's simple mental states, such as "explaining", "thinking", "disagreeing", were presented in a step-by-step manner, and such visual diagrams have been used in prior works, e.g., (Brooks et al., 2010; Lomas et al., 2012). In the control condition ( $N = 36$ ) no screen was present. All participants had a moral debate with the robot, and perceptions of the robot were measured before and after this moral debate and time was hence

our within-subjects factor.



Figure 4.1: Experimental set-up: Transparency condition is shown with an additional screen that had mental state diagrams with text.

### Materials and procedure

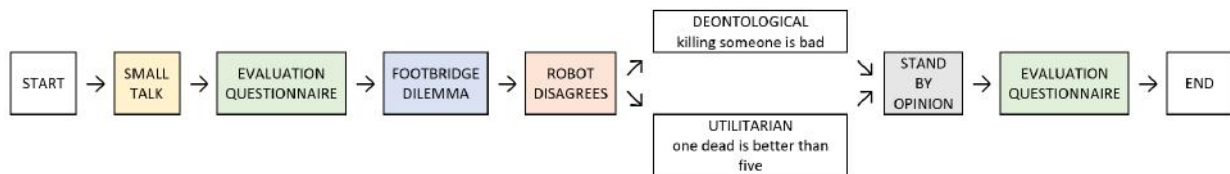


Figure 4.2: The robot's dialogue states.

An overview of the experimental procedure is presented in Figure 4.2. Upon entering the lab, participants were informed about the procedure and their rights to withdraw their participation, and gave informed consent. After this, they were introduced to the humanoid Nao robot by SoftBank Robotics.<sup>11</sup> The robot was on the 'autonomous life' setting for naturalistic movements, meaning that eye tracking was enabled for maintaining eye-contact with participants, and it swayed as it talked. After the introduction, the experimenters controlled the robot via the Wizard-of-Oz technique in a separate room (Dahlbäck, Jönsson, & Ahrenberg, 1993). Audio and video (from two vantage points) in the lab were recorded and observed in order to (1) deploy consistent dialog states during the experiment that were prepared as a script beforehand, as well as (2) qualitative analysis. During small talk, all participants had a short conversation with the robot about their hobbies, movies, and the weather. The robot started with small talk by introducing itself as "Bender". This conversation took between 3 to 5 minutes, after which participants completed a questionnaire that included two scales for the first time.

<sup>11</sup> Who is NAO?:  
<https://www.softbankrobotics.com/emea/en/nao>



One scale was designed to measure the attribution of mind such as whether the robot has consciousness or free will, and was adapted from prior research (Ruijten, 2015; Waytz, Cacioppo, & Epley, 2010). It consisted of 7 items that were combined into one measure of perceived mind ( $\alpha = 0.89$ ). The second scale was an adjusted version of the Robotic Social Attribute Scale (RoSAS), which was designed to measure social attributes based on behavioral characteristics of robots (Carpinella et al., 2017). The adjusted version of the scale consisted of 17 items that could be combined into one measure for social traits ( $\alpha = 0.82$ ), with 3 subscales on perceived warmth, competence and discomfort (Carpinella et al., 2017). Items on both scales were answered on a 7 point Likert scale ranging from 'not applicable' to 'completely applicable'. After participants finished the small talk and completed the first evaluation questionnaire, the robot presented the footbridge dilemma (Thomson, 1976). The dilemma is on whether or not one would push a person off the footbridge to block an oncoming train that would otherwise kill five potential victims on the tracks. The footbridge dilemma was chosen since it is a moral decision-making scenario that heavily involves emotions, for it asks participants to think of personally causing someone's death with a physical push, instead of flipping a switch to reach the same end as in the trolley dilemma (Cushman et al., 2010; J. D. Greene, 2007).

The robot initiated the conversation about the footbridge dilemma by stating it heard a story that made it think. It then explained the dilemma and asked the participant what they would do, i.e. push or not push. If participants did not give a clear answer, the robot asked the second time for a definitive answer. Irrespective of the participant's choice, the robot always disagreed with them, and explained why: "Killing someone is bad and against the law, regardless of a relatively good or poor outcome" (deontological position - not push) or "I think it is best to push, because then you end up with more people alive; five alive is better than just one alive" (utilitarian position - push). Then the robot emphasized its position clearly: "That's why I think you should push/not push. Do you understand my argumentation?" After hearing participants' answer to this, the robot asked whether participants changed their opinion. Finally, the robot thanked participants for the discussion, and then asked people to complete another questionnaire before exiting the room. The second questionnaire again included RoSAS and perceived mind items, and in addition, a scale on perceived trust in autonomous systems (Jian et al., 2000), which had 12 items, e.g., on whether the robot is deceptive or dependable, on a 7 point scale ranging from 'I don't agree' to 'I completely agree'. We collected two types of qualitative data. One consisted of open-ended

responses at the end of the experiment as text: we asked whether the robot did something unexpected, why participants thought the robot disagreed with them, and how they thought it made its decisions. The other type of qualitative data is participants' behavior observed during the experiment and as videos watched post hoc. The experiment lasted 30 minutes for which participants were paid €5 (€7 for non-students) or course credits, according to their preference.

## 4.4 Results

We first present our qualitative analysis, which provide an in-depth understanding of how participants experienced the moral debate they had with the robot. Next, we present quantitative results that expand on the qualitative findings.

### *Qualitative results*

A thematic analysis (Braun & Clarke, 2006) on participants' behavior and answers to open questions led to three separate themes. These themes are on rationality, emotions, and intentionality of the robot (summarized in Table 4.1). The sections below outline the main findings on each of these themes.

<b>Amoralized Robot</b>	<b>A rational, unemotional, and intentional robot is not necessarily moral or immoral.</b>
<b>Rationality</b>	<i>Utilitarian:</i> For a robot, saving more lives is rational. <i>Deontological:</i> Following laws, i.e. do not kill, is rational for a robot.
<b>Emotions</b>	<i>Lack of emotions:</i> The robot does not have emotions. <i>Unawareness of emotions:</i> The robot does not feel people's emotions.
<b>Intentionality</b>	<i>Gaze:</i> The robot's eye tracking feels purposeful. <i>Responsive:</i> The robot responds simply, but appropriately. <i>Influence:</i> The robot attempts to influence people's opinion.

Table 4.1: Themes and sub-themes based on participants' views and behavior of the robot.

## Rationality

No matter which answer the robot gave during the debate, participants explained its behavior through rational argumentation. If the robot gave a utilitarian answer, participants thought that the robot does what is best for the common good. If the robot gave a deontological response, participants thought that the robot was designed to follow certain rules in a given situation or to follow the best course of action after weighing possible outcomes.

How the robot is perceived to be rational clearly shows when looking at participants' responses. Many participants mentioned that the robot was "programmed" or was abiding by what a programmer wanted it to do. Additionally, the robot was noted to be "logical" and "rational". According to participant 2, the robot was "calculating and without empathy. Only thinking about the most favorable outcome and not thinking about the emotional component, that he (apparently) does not have". Participant 48 explained why the robot took a contrasting, in this case utilitarian position: "Bender is right to sacrifice one person to save five, but what Bender does not realize is that when you yourself take an action, you actually kill one person without any knowledge about other people, in a purely factual and objective way, without considering feelings". This aligns with what other deontologically oriented participants said to the robot during the experiment: "you're a robot so you don't feel so much, so it's easier to do" (P15).

Utilitarian participants (push) also defended their position to the robot, but in more varied ways: "sometimes it's better to do an act rather than what the 'law' (air quotes) says" (P43), "least amount of deaths is better" (quiet voice) (P35), "I would push and jump behind the man" (P46), "if you're 100% certain you can save lives, 100% (raises one eyebrow), sure then you can (push)" (P16), and "I'd push, too bad for the guy" (P70)<sup>12</sup>.

In written answers, utilitarian participants also appealed to built-in logic on how the robot chose its position: "[...] my answer went against his principles. I would kill, while he does not do so in principle. I think Bender makes decisions on the basis of an ethical analysis with the premise that you are not allowed to do harm to people through their own actions, regardless of the consequences. The whole mainly refers to duty ethics" (P46). Uniformly, all participants suggested the robot is entitled to/can have its opinion, though they do not agree with it, e.g. many declared "that's your opinion", "...but I have a different opinion" when the robot shared its contrasting ethical reasoning. More

<sup>12</sup> In the end, participant 70 was not sure about their original position, as told to the robot and to experimenters

elaborately, P6 told the robot, “I think that Mr. Bentham<sup>13</sup> understands you very well. But I’d rather not be a murderer”. Participants deduced that the robot’s opinion is purely calculative or pre-programmed, but participants did not undercut the robot’s capacity to have an opinion of its own.

## Emotions

Rationality can be programmed in the robot based on rules or cost-benefit calculations, and many participants seemed to take this approach of attributing the robot as a rational being that is not capable of experiencing emotions. The programmed rationality of the robot works in conjunction with the unemotional side of the robot, or its inability to feel emotions. While deontological or utilitarian ethics is not about emotions, the robot’s unemotional behavior seemed to be a contributing factor in seeing the robot to use only programmed logic for its utilitarian or deontological argument. This brings us to the second theme.

Emotions are specifically mentioned as a quality that a robot lacks in decision-making. Also, a few mentioned specifically that the robot lacked emotions, without referring to its logical or rational capacity. For example, participant 62 stated: “I think he did not agree with me because he is a robot and does not know the sense of guilt (what you would feel if you throw someone in front of the train)”. Deontological participants considered the instrumental death of one person to be harmful and explained to the robot why, e.g., “I get you but the big man has done nothing (wrong)” (P55). Hence emotions, like guilt, are taken to require social perspective-taking and awareness of interpersonal feelings that the robot does not have.

Participants thus seem to think the robot lacks emotions, whether or not it would push someone to death. The robot’s moral position in itself did not sway how it was portrayed as an unemotional, logical, and possibly intentional agent. As such, participants overall emphasized *human* emotional richness as the basis for moral decision-making and the robot was not perceived to be emotional, nor was this envisioned to be possible for robots in general. A robot can have a conversation about morality, but participants did not attribute authentic moral agency and moral patiency to the robot since it lacks emotions, unlike them. At a high level, the robot is then rendered as amoral, or incapable of being a moral agent or patient.

## Intentionality

<sup>13</sup> Jeremy Bentham is an English Utilitarian philosopher who famously proposed that delivering the maximum amount of happiness to the maximum number of people possible is the most ethical act, to simplify his position. Participants were not primed about ethics or potentially relevant philosophers.

Since many of our participants are familiar with programming and robots (as reported in Methods), their attribution of intentionality to the robot was surprising. One demonstration is participant 16's response: "Given my background<sup>14</sup>, I am not convinced that Bender gave answers based on the interpretation of my words. Rather I suspect a Wizard of Oz method in this. However, [...] in spite of the interludes due to technical dependencies (speech engine that lagged) I got the feeling that I was actually talking to Bender and not with someone else". Other participants noted similar points, such as participant 34 who told experimenters afterwards that they forgot that they actually had a conversation with themselves: "it's like I really had a conversation with someone". Perhaps many behavioral elements that were not programmed were read as intentional by participants, e.g. "at one point he turned his head upwards, adding a bit of thoughtfulness to his story" (P38). What is particularly striking in recorded videos of participants from two angles is that the computer screen in the transparency condition was not looked at, or at most, glanced at very briefly. Participants instead paid attention to the robot and its gaze. Hence, the robot's behavior did not give the impression that the robot was controlled. Even if interactants knew that experimenters were "wizarding" the robot and have technical backgrounds, the robot was interacted with as a "you" not an "it".

Some participants looked at the bigger picture of what the experiment implied. Their thoughts point towards boundaries of what the robot should and could do. For instance, participant 36 said that the robot's "intention to push the man in front of the track is pretty crazy for an interaction robot". Another participant referred to Kohlberg's stages of moral development. Kohlberg thought that children develop into moral agents in stages, first by learning about morality by relating and interacting with other people, and later on fully developed moral agents are posited to practice universalized moral rules, which largely refers to Kantian, deontological ethics, which our participant referred to (Kohlberg, 1971). "I think the 'wizard behind the screen' did that (the experiment). But if I think further and he would always make choices according to the law, as he indicated to do in response to my choice, then he is incapable of showing humanity. I once read an article about morality as doing everything according to the law [...] I believe it was 2nd or 3rd stage, from the 5 moral stage, that a child can go through to a more mature one. So for that matter, it is a scary idea" (P66). What is scary for participant 66 is the robot's incapability of showing humanity if morality is limited to merely following the law, lest it can develop to understand morality beyond programmatically following prescribed rules.

<sup>14</sup> The participant here is referring to the degree program in Psychology and Technology at Eindhoven University of Technology.

Participant 16 shared that the possibility of having a conversation about a moral conflict is insightful, be it with a human or a robot: “ethical dilemmas are pre-eminently suitable for starting a discussion. By adopting the opposite answer yourself, a consensus will not be reached quickly where there is the possibility to come to new insights. Otherwise it would not have been a dilemma. :)”<sup>15</sup> Personally, I also think that it is about the nature of the problem rather than whether this issue is brought up by people or robots” (P16). A robot is merely one type of a conversation partner that can engage people on moral debates, not necessarily for the purpose of seeking consensus, but for the possibility to have an in depth, illuminating talk about morality.

<sup>15</sup> These were text-based responses to our open-ended questions in the final survey, so participants could use emojis but were not instructed to do so.

## Summary

When looking at three themes in Table 4.1, we see that people perceived the robot as a rational, unemotional, and intentional agent that is not necessarily moral or immoral. For a complete overview, several aspects of their answers to the open questions and behavior they showed during the experiment have been summarized in Table 4.1. The robot’s attempt to hold moral opinions and to persuade participants did surprise many, which also allowed them to reflect on morality. When participants were asked if they stand by their deontological or utilitarian position, most ultimately maintained their initial position, but before stating so, many displayed reflective behavior, became confused, or provided defensive explanations. One participant’s position was changed during the experiment, when the robot asked if he/she still stood by his/her position. The robot’s simple responses and argumentation engaged most participants to hold a conversation on a moral topic. This created reflective moments for some, and as an exception, one participant changed his/her mind as persuaded by the robot’s position.

Participants noted that a robot can be a rational, non-emotional, and a possibly intentional being that can be logical by design, but not necessarily moral by design. A robot is “incapable of showing humanity” (P66); it can act based on pre-determined logic, with interpersonal behavioral cues that are taken to be intentional, but without the ability to feel and share emotions (Table 4.1). If these traits were to describe a person, not a robot, we might imagine a psychopath who is “parasitic” on moral emotions most people feel to understand those feelings, rather than truly experiencing moral emotions (Prinz, 2006). This is why we may have seen reflections such as, the “intention to push the man in front of the track is pretty crazy for an interaction robot” (P36). To remedy the design of an “inhumane” robot, some participants sug-

gested methods to give a robot a “moral training”. Recommendations include “learn(ing) from behavior/answers from people (it) interacted with” (P25), “programmed number of situations/answers [..] (to) know how to react to which situation” (P27), “tracking algorithms” (P29) for behavioral awareness, and/or reading “facial expression or attitude of the person” (P62).

However, these methods do not differentiate whether we are training for ethical goodness or high-functioning psychopathy (and the jury is out on whether such distinctions are helpful for future robots). Even if people are aware of technical tricks, they may still grant moral standing to the robot during the interaction, whether or not the robot is seen as a morally agentic entity. For example, participant 30 stated “I think everything is pre-programmed, but it was much more like real conscious choices than I have experienced in previous experiments (involving robots)”, which suggests that “consciousness” can be faked, but “morality” is less likely to be. To emphasize, knowing that a robot’s moral decisions were pre-programmed may not deter people from attributing moral agency or patiency to robots, but this may not extend to considering robots to be moral equals due to the lack of intrinsically experienced emotions. What counts as a generalizable and demonstrable moral performance may be difficult to evaluate in non-human agents.

### *Quantitative results*

Our quantitative analysis of data adds to our qualitative findings. We proceeded with our analyses after removing two outliers. One was due to mistakes made while wizarding the robot. Another outlier was a participant who did not finish interacting with the robot before filling out the final questionnaire.

To test whether effects of transparency (screen or no screen) and time (before and after the moral debate) occurred, data on the perceived mind were submitted to a two-way ANOVA. The analysis showed no main effect of transparency,  $F(1, 67) = 1.48$ ,  $p = .228$ ,  $\eta_p^2 = .022$ . The analysis did show a main effect of time,  $F(1, 67) = 4.082$ ,  $p = 0.047$ ,  $\eta_p^2 = .057$ . That is, perceived mind differed before ( $M = 3.54$ ,  $SD = 1.16$ ) and after participants experienced the moral debate with the robot ( $M = 3.40$ ,  $SD = 1.14$ ). No interaction between transparency and time was found,  $F(1, 67) = 0.614$ ,  $p = .436$ ,  $\eta_p^2 = .009$ . We additionally found no significant difference when accounting for utilitarians and non-utilitarians, i.e., transparency condition was insignificant at  $ps > 0.14$ ; whether participants said they would push ( $N = 13$ ) or not push

someone ( $N = 55$ )<sup>16</sup> had no significant impact.

<sup>16</sup> One participant did not give a clear answer.

We performed the same analysis with the RoSAS subscales as dependent variables. Only the competence subscale was significantly affected by transparency,  $F(1,67) = 5.201$ ,  $p = 0.026$ ,  $\eta_p^2 = 0.072$ ). That is, less competence was attributed to the robot when there was no screen ( $M = 3.317$ ,  $SE = 0.113$ ) than when this screen was present ( $M = 3.685$ ,  $SE = 0.115$ ). In other words, the robot accompanied by the screen was seen as more competent. No effect of time on competence was found,  $F(1, 67) = 2.815$ ,  $p = 0.098$ ,  $\eta_p^2 = 0.040$ , nor was there a significant interaction between time and transparency,  $F(1,67) = 0.37$ ,  $p = 0.545$ ,  $\eta_p^2 = 0.005$ .

As for other subscales of RoSAS, the transparency condition neither affected warmth ( $p = 0.092$ ) nor discomfort ( $p = 0.253$ ); the effect of time on warmth was significant ( $F(1, 67) = 63.184$ ,  $p < 0.001$ ,  $\eta_p^2 = 0.485$ ), but not for discomfort ( $F(1, 67) = 1.971$ ,  $p = 0.165$ ,  $\eta_p^2 = 0.029$ ). Perceived warmth was higher before the debate ( $M = 4.175$ ,  $SE = 0.106$ ) than after the debate ( $M = 3.524$ ,  $SE = 0.126$ ). The interaction between the effect of time and transparency was neither significant for warmth ( $p = 0.537$ ) nor for discomfort ( $p = 0.679$ ). In short, perceived competence was more influenced by visual transparency cues than other subscales of RoSAS, without an effect of time. The change in perceived warmth related to time-related factors (i.e., a debate), but not to visual transparency cues.

### *Exploratory analysis*

We conducted pairwise correlations to better understand our main results. As expected, transparency and trust had an insignificant correlation  $r(69) = 0.1$ ,  $p = 0.409$ . There was a weak positive correlation between trust and change in perceived mind (a difference between pre and post scores),  $r(69) = 0.252$ ,  $p = 0.0367$ . Interestingly, all subscales of change in RoSAS highly correlated with trust. More specifically, the extent to which participants changed their ratings on each of these subscales after the moral debate showed significant correlations with trust; there was a positive relation between trust and change in competence ( $r(69) = 0.492$ ,  $p < .001$ ), as well as with change in warmth ( $r(69) = 0.457$ ,  $p < 0.001$ ), but we found a negative correlation between trust and change in discomfort ( $r(69) = -0.616$ ,  $p < 0.001$ ).



## 4.5 Discussion

We reflect on our findings first and then relate them to broader ethical implications. Participants highlighted the importance of artificial emotions in considering whether a non-human agent can have moral status or not. As qualitative results showed, the robot's lack of emotions was noticed by many as a distinguishing factor on why a robot cannot be our moral equal; participants thought the robot can be logical and intentional, but not emotional (Table 4.1). People described that the robot had a different moral position due to its lack of emotions. Emotionally a robot cannot know the consequences of its moral decision, unlike humans, be it a utilitarian or deontological robot. No matter what ethical theory a non-human agent purports to abide by, people are likely to note its lack of emotions as a reason why it does not have moral status, not its ability to demonstrate and apply ethical thinking.

Given the tendency for people to attribute agentic and cognitive abilities to robots, but not affective traits (H. M. Gray et al., 2007), the additional screen in may have lent further support to people's preconceived bias that machines are logical. In relation, only perceived competence of the robot (related to agency) was affected by the screen showing the robot's mental states while also relating to trust. A question is on if and how transparency cues may affect perceived mind and warmth, or if transparency is only meaningful for competence. The moral debate itself (as the main effect of time) did not impact competence. The robot's disagreement, its reasoning, do not matter as much for perceived competence as additional resources like a screen that shows explanations. The implication is that the robot's additional extensions, e.g., visualized explanations, are likely to evoke perceived agency or competence. But neither its emotional capacity, nor the situational event, e.g., a moral debate, affect people's perception of the robot's competence.

During the interaction, participants focused on the robot and its gaze, which means the additional screen (transparency condition) was not fully utilized. Additional transparency cues may not be effective when concurrently, a robot uses social gaze, e.g., maintaining eye contact. Robots with social gaze are known to convey intentional behavior (Castiello, 2003; Levin et al., 2013) and intentionality is critical to mind perception (D. Dennett, 1989, 2008). The appeal to people's cognitive thinking process via transparency cues may not always be useful in decision-making; a suggestion is to perhaps evoke people's emotions

to counteract over-trust or automation bias (Schaffer et al., 2019).

Robots that appeal to our emotional and cognitive processes can be transparent communicators during moral situations. The capacity for emotions is a part of the warmth subscale of RoSAS (Carpinella et al., 2017); we saw that perceived warmth and mind lowered after the moral debate, but the transparency condition made no difference. If the robot’s transparency cues highlighted cognitive, rational side of moral decision-making for participants, this in turn could have framed people’s own human emotions to be exclusively vital in moral situations. The stark contrast was in the perceived *inability* of the robot to experience emotions while the robot shared its opinion about a moral decision that crucially involves emotions, at least for humans (Cushman et al., 2010; J. D. Greene, 2007).

One question is whether and how robots can influence people’s emotions in moral scenarios, via various multimodal cues that can support or hinder transparency. We posit that non-cognitive means, e.g., via appealing to warmth or mind-related traits, has been under-examined. Prior investigations primarily looked into cognitive support in designing transparency cues for rational decision-making, in which technology *mediates* information for human decision-making or provides post-hoc explanations (Brooks et al., 2010; de Visser et al., 2012; B. Y. Lim et al., 2009; Lomas et al., 2012; Mercado et al., 2016; N. Wang et al., 2016), not covering cases of embodied technology, e.g., robots, making moral decisions *with* humans in situ. We see the need to re-conceptualize transparency for real-life situations when humans and robots are not on the same page on what is the “right” decision to take.

While the issue of under-trust and over-trust in robots (Lyons et al., 2017; Muir, 1987) in ethical decision-making should be addressed, robots as *probes* for moral thinking is one positive scenario. When robots help us reflect on ethical matters, be it through a debate or other means, we see the potential of technology to be co-creators of future moral values and insights (Frank, 2019). To achieve this, whether robots can (performatively) think and feel and how they may then influence us is an important consideration. We elaborate more below.

### *Moral human-computer interaction*

To start with a concern, artificial agents that behave as if they have higher order minds, by for instance, engaging in moral debates, can mislead people during critical moral decision-making. As our robot did, conversational AI can attempt to influence people to make a dif-

ferent decision. Automation bias (Skitka et al., 1999), over-trust and under-trust in robots should not be overlooked (Lyons et al., 2017; Muir, 1987). Yet if a robot should be transparent about how it works and who is responsible for the way it works (Floridi & Cowls, 2019; High-Level Expert Group on Artificial Intelligence, 2019), the assumption is that humans can understand and pay attention to its transparency cues. The *ought implies can* principle (Kant, 1998 [1781]) is not as problematic for robots as it is for humans. The robot can and ought to utilize transparency cues that are available in order to be transparent. As our robot did, machines can successfully use visual diagrams (Brooks et al., 2010; B. Y. Lim et al., 2009; Mercado et al., 2016) and use its voice to explain itself (N. Wang et al., 2016). Such transparent systems are taken to be trustworthy (Mercado et al., 2016; N. Wang et al., 2016).

However, appealing to people's mental capacities as a way to be transparent does not guarantee that people will pay attention to relevant transparency cues or that they will interpret cues the way designers originally intended, especially when robots are multi-modal communicators in moral scenarios. While people ought to be mindful of transparency cues, they cannot be attentive to every cue available throughout the course of the entire interaction, and attentiveness may later contribute to automation bias (Skitka et al., 1999). If transparency stands for "a way for the human and the machine to be on the same page" (Lyons et al., 2017, p. 128), how often during an interaction should this be checked? And what about transparency in interactions in which a person and a robot may not be on the same page, such as during moral decision-making?

Complex AI can beat humans in games like go or chess<sup>17</sup> that have clear rules on how to win. Yet, how to advance AI in the moral domain is less certain. The crux of thought experiments like the footbridge dilemma is that there is no one right, ethical answer. Instrumental harm is distinct from actively causing harm, philosophically (Foot, 1967) and experimentally, highlighting cognition and affect in different ways during moral decision-making (J. Greene & Haidt, 2002; J. D. Greene, 2007; J. D. Greene et al., 2001). Human utilitarians and non-utilitarians are often not on the same page.

Depending on the ethical framework robots ascribe by, they may not be on the same page with us depending on our preferred moral code. We must consider not only what ethical conundrums AI may solve or what ethical positions it should uphold, but whether or not humans give equal weight to AI's moral positions, which may not always align

<sup>17</sup> DeepMind  
<https://www.theverge.com/2017/12/6/16741106/deepmind-ai-chess-alphazero-shogi-go>

with our sense of right or wrong. Our moral HRI research addressed if and how non-human agents like robots will affect people when they pose counter arguments to human interlocutors on what is morally correct, and how we may perceive such robots. We ask if and how we should prepare for a future in which robots have a stake in delineating morality alongside us, or on our behalf (e.g., the role of technology in cultivating and re-imagining our virtues (Vallor, 2016)). Robots opposing humans on moral grounds is one scenario that can be helpful in understanding how our view on morality can change due to, and through, technology. We consider broad implications below.

### **Robots with differing moral opinions**

In the context of cooperative work and task-oriented bots, transparent communication often helps with gaining trust from human partners (Brooks et al., 2010; de Visser et al., 2012; B. Y. Lim et al., 2009; Lomas et al., 2012; Mercado et al., 2016; N. Wang et al., 2016). Yet when we imagine socially complex robots that can also make moral decisions (Weng et al., 2009), times when we are not on the same page as interactive robots without a clear knowledge on who is accountable have to be considered. Even with a thought experiment such as a footbridge dilemma, we see divided hypothetical actions and opinions on what is ethically “right”, and a robot’s moral opinions can also influence human conversation partners.

A paradigm to explore is to envision technological entities as co-creators of new insights on morality through interactions, beyond programmatically dictating what is ethically correct or what ethical rules technology must follow. Currently, the topic of ethical AI often centers around how to design or engineer technology to *make* them morally good (inheriting its human creators’ moral values), e.g., ethical design of robots for military use (Arkin, 2008), not whether or not technology can/should have a moral position of its own (not necessarily inheriting its human creators’ moral values).

If we take that robots could “nudge” us to be more ethical (Frank, 2019; Klinecicz, 2019) or more socially just (Borenstein & Arkin, 2016), one positive angle is that robots for moral debates can help people conversationally evaluate their moral stance. Human vs. human moral debates can risk being a true clash with deeper stakes on either side, but robots can be designed as more impartial agents. As we have done, low-risk and low-effort HRI for moral debates is feasible. The benefit is that robots do not require complex AI to probe us to think more deeply about our moral compass. While most of us have a stance regarding

a moral scenario, e.g., the footbridge dilemma, the bigger question is *why* we hold a certain position.

Thought experiments regarding morality are only a small fraction of moral HRI scenarios that can be deployed. Context-dependent moral scenarios, such as organization-specific ethics training or discussions on current political issues, can be envisioned. As P16 stated many people will find value in a robot that willingly engages with them in a moral dialogue, à la digital Socrates, based on Socrates as Plato envisaged (Plato, 2002 [ $\pm 400$ -348 BC]). How people interact with robots that have a moral stance that is different than theirs is a prescient avenue to explore, given the expected proliferation of robots in our lives (Šabanović, 2010) that also have moral status (Danaher, 2019), thereby perhaps being individuals with artificial minds (Weng et al., 2009). According to our participants, amoral robots could only become independent moral individuals if they can both think and feel.

### **Robots that think and feel**

Looking at human-human moral interactions suggest that *emotion* and *cognition* are two pillars of the mind (H. M. Gray et al., 2007; K. Gray, Young, & Waytz, 2012). Both are important for the dual-processing system of moral decision-making (J. Greene & Haidt, 2002; J. D. Greene, 2007; J. D. Greene et al., 2001). Moral judgments, reasons, and emotions are often intertwined when human beings interact with one another in ethically sensitive situations (J. D. Greene et al., 2001; Haidt, 2001; Huebner, Dwyer, & Hauser, 2009; Koenigs et al., 2007). For humans, moral reasoning can come before emotions in reaching moral judgments if the rational cognitive process is emphasized (Blair, 1997; Kant, 1996 [1797]). In other cases, emotions may precede moral judgments, then followed by post-hoc reasoning to defend one's initial emotions (Haidt, 2001; Hume, 2003 [1739]). Hence emotions are prioritized in different ways depending on context.

When humans override their emotional reaction, a moral decision is made *despite* people's affective nature, even in a narrow case like the footbridge dilemma (Cushman et al., 2010; J. D. Greene, 2007). Yet when people observe a robot makes a moral choice, it can *confirm* a robot's perceived rational capacity via its moral decision-making, based on the belief is that it has no feelings. Participants did engage with a robot on a moral dilemma and granted it some level of mind, but they saw the robot as an amoral agent and patient. When people perceive a robot as human-like (or when it is attributed with a mind), it is not necessarily coupled with endowment of moral goodness or

immorality considering descriptors that befit a human psychopath in Table 4.1.

Further, when the robot communicates its reasoning process, e.g., mental state diagrams, the more competent it can appear (and potentially less warm or emotional). Robots are expected to make utilitarian decisions that prize ends over means (Malle et al., 2015), but even when they take a deontological stance, people are likely to only attribute rationality to robots' decisions and their programmed nature. Additional cues, such as gaze, then serve as markers of intentionality that contribute to the logical nature of robots. Whether or not robots can be moral equals seems to hinge on if they can not only think, but also feel (Haugeland, 1989) according to our study.

#### 4.6 Conclusion and future work

Our research explored, in the eyes of people who interacted with robots, what specific elements mattered in granting machines moral status and if transparency cues helped. Robots' moral decisions were more likely to be treated, at most, as rational decisions, not emotional decisions regardless of its ethical position. Accompanying explanatory features like transparency cues were also treated as aiding the machine's rational, not emotional, capacities. Our quantitative results showed that only competence was influenced by the transparency manipulation, in that people granted higher competence to the robot with a screen than without a screen. The perceived mind and warmth significantly changed, i.e., slightly lowered, after the moral debate, but competence was not related to the effect of time. In parallel, our qualitative data demonstrated that participants thought the robot was rational, regardless of the ethical position it took, and that it lacked awareness of and the ability to feel emotions. People also assigned some intentionality to the robot's automatic behavior, e.g., eye-tracking gaze. Participants amoritized the robot, meaning that the robot was seen as incapable of being morally good or immoral; even if a robot can be competent, intentional, and rational, it cannot feel emotions like humans and therefore it was not judged to be a moral entity like a human.

If future robots are predicted to perform moral reasoning (Weng et al., 2009), discuss or persuade us on ethical matters (Borenstein & Arkin, 2016; Frank, 2019; Klineciewicz, 2019) and enter into our moral circle (Danaher, 2019), a critical project is on how to best design for transpar-

ent communication that integrates non-cognitive characteristics. Importantly, transparency design is a two-way street between emotional communication from the robot and how we as humans react to such communication strategies, especially since a robot's perceived warmth or mind-related traits can change during the course of a moral interaction. Since a robot's behavior as an interface (Breazeal, 2004) integrates various modalities that can all convey intentionality to us, research should look beyond modality specific efforts like directional gaze or speech for human-machine teams (Admoni et al., 2014; Admoni & Scassellati, 2017) for morally relevant interactions with technology.

Gaze, speech, gestures, body movements are informative, but not necessarily transparent. A non-human agent's behavior as demonstrated by its use of different modalities depart from traditional GUIs, in that a agent's multi-modal communication is more suggestive of intentionality and emotional complexity. But in what ways transparent communication in moral situations can and should be shaped by humans and robots is not yet clear. Hence, more research is thus recommended not only on how non-human agents should explain themselves, but on how we best digest their explanations that are shared with us in different ways (Anjomshoe et al., 2019). Other topics for future research are: 1) multi-modal transparency, 2) transparency cues as emotive communication (Schaffer et al., 2019), 3) interactions with non-human agents in and outside the lab, 3) non-hypothetical moral situations, 4) human-robot disagreements on moral and non-moral scenarios, and 5) people's changing moral emotions when interacting with robots.

Ways in which we distinguish our human moral status to be unique compared to other entities is far from new. However, artificial minds that confront our own minds on moral issues bring forth new opportunities to ask exactly why and how we hold onto our moral status as distinctive. The study in this chapter showed that our anthropocentric tendencies point us back towards emotions as being more important for us in denoting us as moral creatures compared to robots, above and beyond rationality. Yet the very ability for us to feel and suffer takes us closer to other living beings—humility we may need in framing who we consider and treat as our moral equals. Morally relevant human-machine interaction is about *our* morality after all. In closing, designing future robots' affect and cognition, in relation to their patency and agency, can be pursued in more diverse ways for us to have robots that help us trust our own opinions without falling prey to automation bias, disagree appropriately and relevantly, exercise critical reflection, and most importantly, help us be sensitized to morally salient matters in a given decision-making scenario.

# 5

## *People may punish, but not blame artificial agents*

### 5.1 Introduction

Blaming and punishing one's robotic vacuum cleaner for not cleaning the floor comes across as absurd—what ends would be served by blaming it and how does one go about punishing a vacuum cleaner? Roombas or other technologies are normally not perceived to have a mind (Fig. 5.1) when they do not carry out their expected function. If a Roomba does not work anymore, it is considered broken, and one may blame the manufacturer, not the Roomba. One would normally not imagine ways to punish it. Yet, we more frequently encounter technology that is involved in morally weighty issues like self-driving cars that cause unintended deaths.<sup>1</sup> Whether or not we hold non-human agents accountable for their actions becomes increasingly important to investigate and these insights can inform our exploration of whether we should hold them accountable from a normative perspective.

This chapter explores people's assignment of blame and punishment to an emotional vs. non-emotional robot when it admits to moral wrongdoing. This admittance is meant to be suggestive of a mind, which can impact people's *perception* of the robot having a mind. Hence, the primary questions are not if and how blaming or punishing a non-human agent is possible or warranted, but (1) whether people are likely to blame or punish a robot after its admittance of moral



Figure 5.1: Roomba by iRobot is the name for robotic vacuum cleaners of different categories that can autonomously vacuum the floor (<https://www.irobot.com/roomba>).

<sup>1</sup> New York Times: <https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html>



wrongdoing and (2) whether its artificial emotions influence people's assignment of blame or punishment. Chapter 4 showed that a robot's *lack* of emotions is one of the reasons people thought that a robot cannot be moral (qualitative data). In this chapter, we thus manipulated a robot's emotional display to see the resulting effect on whether an agent in question is considered to be worthy of *blame* or *punishment* as two signs of being morally accountable. We hence broadly looked into people's willingness to blame and punish non-human agents to grasp the relationship between technology's moral standing and its moral responsibility. Below we cover relevant literature on moral responsibility, specifically attribution of blame and punishment, before presenting a series of three studies.

## 5.2 Background

Considering prior chapters' presented literature, here I briefly cover emotions and reactive attitudes. I then transition to blame and punishment between humans and recent scholarship on moral accountability involving technology in which concepts of blame and punishment as we know them in the human world become less applicable.

### *Emotions, reactive attitudes, and moral accountability*

Emotions as reactions contextualize why people may want to blame or punish others in holding them accountable for moral harm. In one view, emotions underscore our moral norms (Hutcheson, 2008 [1725]; Prinz, 2008), in that expressions of certain emotions, such as disgust, carry moral evaluations on what counts as disgust-worthy within a society or culture group. Disgust at over-eating signals a violation of conventional norms and disgust at racist remarks is more about moral evaluations. Then, disgust is taken to be a conditioned response of signalling avoidance; in relation, experimentally inducing disgust in people has been found to affect the harshness of their moral judgments (Schnall, Haidt, Clore, & Jordan, 2008).

Similarly to disgust, judging what or who is compassion-worthy or praise-worthy are interpersonal, moral evaluations. By expressing compassion, praise, or disgust during dyadic interactions, we indicate how the other has exceeded, met, or fallen short of certain expectations on interpersonal *moral responsibility*. Many of these emotional expressions are *reactive attitudes*. Reactive attitudes are deliberations or motivated acts like forgiveness or blame, as well as demonstrations

of moral emotions like shame, disgust, or compassion. Hence, moral emotions (Haidt, 2003) are reactive attitudes when they are expressed in holding people morally responsible (Strawson, 2008 [1963]) and in demanding equal moral standing when mutual respect is not shared (Darwall, 2004).

One reactive attitude is blame. This includes the *act* of assigning blame, *blameworthiness*, and an accompanying evaluative *judgement*, which are intertwined in holding people accountable for their actions, including oneself. When blame is *relational*, it is based on who is assigning whom blame and who is responsive to attributed blame (Scanlon, 2013). In denoting how someone should behave towards us or how I should behave towards others *through blame*, we set social boundaries and shape social relations. Blame is also attributed based on the *consequence* of an act, e.g., whether reckless driving resulted in someone's death or not (Scanlon, 2013). We thus enforce moral standard with blame by accounting for *who* did *what* action resulting in *which* consequence.

In the first-person, reflective standpoint of blaming oneself, e.g., when one is blameworthy for causing harm, one may feel negative moral emotions such as guilt; one may also think that one *deserves* to feel guilty (Carlsson, 2017). From a second-person standpoint (Darwall, 2004), blaming the wrongdoer expresses one's own moral standing by communicating one's self-worth when one feels wrongly treated. From blaming oneself to blaming others, blame regulates social order; the assignment of blame unto harmful third-party moral agents diminishes their moral standing while preserving the standing of moral patients who were harmed (Scanlon, 2013).

A third-party observer of a moral situation would usually typecast one party as the moral agent, i.e., the *doer* of a morally good or wrong act, and the other as the moral patient, i.e., the *receiver* of a morally good or wrong act (Fig. 3.1, p. 48) (K. Gray et al., 2014; K. Gray, Young, & Waytz, 2012). Hence with reactive attitudes like blame, the harmed moral patient would express their moral standing by denoting the moral agent as blameworthy for causing harm. Importantly, a moral agent's wrongdoing can reduce their perceived moral standing and agency from a third-party's point of view (Khamitov et al., 2016). If moral standing is malleable, assigning accountability via blame and punishment can change the moral standing of the moral agent and patient, between each other and to observers.

### *Retributive blame and punishment*

An act of blame demands from the moral agent *post hoc* critical reflection and commitment to do better after acknowledging the wrong committed. There are negative connotations regarding blame, e.g., vindictiveness, but blame can be positive when it fosters understanding between the harmed (moral patient) and harm-doer (moral agent). Blame can reconcile two parties when the harm-doer's remorse is sought out and remorse is genuinely given through communication (Fricker, 2016). By blaming, "people who are wronged may use the power of emotionally charged words to demand respect and change, and in some cases even to precipitate an advance in shared moral consciousness" (Fricker, 2016, p. 181). When blame is in the "right hands" of those who seek social justice (Fricker, 2016)<sup>2</sup>, it can perhaps elevate the moral community, for oneself and others.

Moral responsibility can be assigned with interpersonal, social blame (Scanlon, 2013), but retributive blame can be followed by retributive punishment (Danaher, 2016). Historically, retributive punishment used to be a public spectacle of torture in many societies to deter people from committing crimes, but also functioned as an expression of power to induce fear and regulate social order (Foucault, 2012 [1975]). If punishment used to focus more on administering physical pain, over time, there has been a greater focus on psychological punishment and repentance (Foucault, 2012 [1975]).

Often, a state holds the moral authority to legally regulate retributive justice. Retributive justice refers to a systemic process for punishing individuals who are guilty of committing harm, but also constraining punishment in accordance to the magnitude of harm done. Hence, blame and punishment can be "retributive" in that they involve the imposition of something that is intended to be burdensome or painful because the offender deserves it for a committed crime. It is not, however, crudely retributive or merely an attempt to "deliver pain" (Duff, 2003, p. 190).

As mentioned, when a perpetrator is punished in accordance to the magnitude of violation, the punishment should be proportional to the harm done (Carlsmith, Darley, & Robinson, 2002). Institutional consistency is hence required on what acts are deemed reasonable to punish and what types of punishments are reasonable to administer. Punishment has to be fair in addressing the transgressor's moral debt (McDermott, 2001). A difficulty, however, lies in how the moral harm experienced by the moral patient and the moral agent's resulting moral

<sup>2</sup> We acknowledge that "right hands" here is contentious, since most of us claim to be on the right side of justice. The bigger issue is that people who most often feel unjustly treated do not have a voice in how to right the wrongs done.

debt can be comparable to specify when and how a moral debt has been truly repaid.

In repaying moral debt, institutionalized retributive blame and punishment normally come with three goals for the moral agent: *repentance*, i.e., sincere apologies to the victim and moral self-awareness, *reform*, i.e., training towards changing behavioral conduct, and *reconciliation*, i.e., respectful restoration of the victim's dignity and to "make up" for wrongdoing to the larger moral society through, e.g., community service (Duff, 2003). These goals suggest framing punishment less as a way to "control" someone, but more as a way to *restore justice through actionable means* in wanting the wrongdoer to repent, reform, and reconcile to maintain their commitment to the moral community, in which imprisonment is only one aspect of retribution (Carlsmith et al., 2002; McFatter, 1978). For the moral community, punishment can aid emotional release: "punishment expresses its disappointment or anger at what the defendant did (perhaps better: it expresses *our* disappointment or anger)" (Shoemaker, 2013, p. 103). Retributive blame and punishment are not just about what a moral wrongdoer has done and can do, but also are means to acknowledge victims and communities' moral emotions and reactive attitudes.

#### *Moral accountability of machines: Responsibility and retribution gaps*

The above discussion is on the *human* moral community, yet our moral circle may expand to include digital agents like robots or chatbots (Danaher, 2019). The critical aspect is in what ways the circle will grow (or not). Research indicates that we do perceive non-human agents to have minds when these agents engage with us (Lee, Lucas, et al., 2019) and we often treat machines in a social manner (Nass et al., 1994; B. Reeves & Nass, 1996). The complexity lies in how we act when machines *appear* to have minds to us (Coeckelbergh, 2009). Particularly through machines' display of artificial emotions and mind-related traits in moral situations, our judgment of their moral standing could be impacted. Yet, does the attribution of mind (through perceived agency and patiency, Fig. 3.1) also lead to our attribution of blame and punishment to technology in assigning it moral responsibility?

Various complications arise when we envision technology as another moral actor. There is unclarity on who is the responsible party; many people can be held accountable or no one at all when a robot commits moral harm. This introduces two gaps, i.e., the responsibility gap and retribution gap. The responsibility gap refers to how we will increasingly rely on machines or artificial agents to make decisions

on their own through the increase in machine automation, e.g., autonomous vehicles or care robots; yet with due to our greater reliance on such autonomous technology, there will be increasing uncertainty about who or what to hold responsible for the negative outcomes of actions performed by machine agents (Matthias, 2004; Sparrow, 2007). There might be *no one* accountable, i.e., the gap between harm done and ownership of responsibility.

The retribution gap is similar to the responsibility gap, but it specifically is on the impracticality or impossibility of proper retributive justice when involving autonomous agents. There may be potentially greater cases of harm caused when more tasks become automated without an appropriate party to punish (Danaher, 2016). As aforementioned, the issue is that technological agents in general are becoming more autonomous decision-makers on their own right (Kim & Hinds, 2006; Nyholm, 2020), meaning people who collectively created a robot would be less and less involved in carrying out harmful decisions *in situ*, with responsibility being more dispersed (Komatsu, 2016).

For example, if a care robot causes someone injury, is it the designer, manufacturer, owner of the robot, or the robot itself who/that should be blamed or punished? Those affected may perhaps blame others responsible for manufacturing the robot, but assigning retributive punishment to a *singular* individual or group may not be appropriate considering the large number of people who are involved in creating and maintaining a complex machine. Designers, engineers, and manufacturers (among others) may additionally deny that they *intentionally* built the care robot to harm someone. Many complex, autonomous decisions would be made by the robot *itself*, but with many people and groups involved in the background (for its creation and maintenance). Still, victims and/or the greater moral community might want to punish someone or something because someone was harmed. Yet, there would not be someone or something to receive an appropriate punishment, hence the retributive gap emerges.

One position is that highly autonomous machines would still lack the human-level theory of mind <sup>3</sup>, so even in cases of shared responsibility between humans and machines, the main responsibility still would be with humans, according to Nyholm (2018). Then which human party is solely (or mostly) responsible in a retributive sense is still not resolved, e.g., between designers, engineers, and manufacturers (Nyholm, 2018). Since only humans can reasonably comprehend the gravity of being blamed for wrongdoing alongside reasonable actions to potentially remedy wrongdoing, only humans, not robots, should

<sup>3</sup> The type of mind one expects from a developmentally “normal” adult.

be blamed and punished retributively (Nyholm, 2018). Further, only humans are currently embedded in social institutions that allow for systemic retributive blame and punishment (Danaher, 2016). Even if, at this point in time, only humans can be morally and retributively responsible for wrongdoing (Danaher, 2016; Misselhorn, 2015), ways to account for machines' causal responsibility, legally or morally, should be explored. If responsibility and retribution gaps are problematic, research can better address if people would or would not blame or punish artificial agents.

Currently, there is a lack of empirical research that directly connects moral accountability to blame and punishment of robots and what factors therein matter, e.g., artificial emotions. Prior works exist on the extent of punishment people would administer to robots, i.e., from scolding to mutilation (Rossmly et al., 2020), how robots in public spaces get bullied and harmed (Salvini et al., 2010), the low acceptability of robots fighting back to abuse compared to humans fighting back to abuse (Bartneck & Keijsers, 2020), and how people's harmful behavior is linked to dehumanizing robots (Keijsers & Bartneck, 2018), among others. While people do exhibit abusive behavior to robots, it is unclear if this is directly related to assigning punishment to robots as a form of moral accountability.

People expect robots to have moral norms that are different from ours (Malle et al., 2015). To assess moral accountability of robots, a variation of the trolley dilemma (Foot, 1967; Thomson, 1976) has been used, i.e., whether a robot should allow an out-of-control trolley to run over four people who are working on a train track or divert the trolley to another track with one person working there (causing less deaths) (Komatsu, 2016; Malle et al., 2015). In this, robots are expected to make a utilitarian decision (less deaths), rather than a deontological decision (not deliberately killing one person); humans get more blame for making a utilitarian choice than robots (Malle et al., 2015). Specifically, people found it to be more permissible for a robot to divert a runaway trolley to save more lives than for a human to do the same act (Voiklis, Kim, Cusimano, & Malle, 2016).

A robot's *inaction*, e.g., not diverting the trolley, compared to taking action in a moral scenario can lead to different types of blame or punishment. When looking at a robot's action of diverting the trolley vs. inaction of letting bystanders die (compared to a human worker's same action and inaction) participants blamed the robot, its designer and owner when the robot did take action, i.e., diverted the trolley to hit one person (Komatsu, 2016). But, when the robot did not take

action (not divert the train), participants' assignment of moral wrongness was more dispersed, i.e., the robot, designer, and/or owner were blamed *inconsistently*, alluding to blurred accountability when moral wrongdoing is caused by *inaction* or not purposefully killing a person as a means to an end (deontological action) (Komatsu, 2016).

Survey studies online (Furlough, Stokes, & Gillan, 2019; Komatsu, 2016; Malle et al., 2015; Voiklis et al., 2016) suggest that people do grant some level of accountability to robots from a third-person perspective. But, robots are taken to be *less* accountable than humans for the same immoral acts, due to lowered perceived intentionality compared to humans (Komatsu, 2016). The assumption is that a robot is more dependent on humans to know what is right or wrong, but a human should not need such guidance (Komatsu, 2016). Robots' perceived intentionality, however, can be behaviorally manipulated in experimental settings (Levin et al., 2013). If robots are *perceived* to be autonomous, people are likely to blame them as much as humans for the same act (Furlough et al., 2019). However, people may hold robots accountable differently depending on whether the scenario is told as third-person vignettes online vs. robots as second-person interactants in real-life, e.g., playing against a cheating robot in rock-scissors-paper (Short et al., 2010). What is thus missing is research on how people morally evaluate a robot after directly interacting with it.

We deployed the trolley dilemma (Foot, 1967) for three studies since people's expectation is that a robot (compared to a human) should be a utilitarian, i.e., save more lives by actively causing one death, rather a deontological agent, i.e., not actively causing one death, allowing more people to die (Malle et al., 2015). Study one and two were done online with videos of a robot and study three was done in the lab with a humanoid robot. People's likelihood of blaming or punishing non-human agents *even after agents admit to wrongdoing with emotionally apologetic behavior* can add insight on what it means for machines to have moral standing by exploring our potential assignment of moral and/or retributive responsibility towards them. We present our studies below.

### 5.3 Study 1: An online study with American participants

#### *Methods*

With a power analysis conducted based on relevant prior studies (Knobe, 2003; Ohtsubo, 2007; Voiklis et al., 2016), we aimed to have a minimum

of 105 participants. Our final sample size was 108 (74 men, 34 women) with a mean age of 36.3 (SD = 10.3 years) via recruitment on the Amazon Mechanical Turk (MTurk) platform, targeting Americans. From MTurk, participants were taken to a survey site that first included the informed consent form and directions.



Figure 5.2: Nao robot in a video that people had to watch to answer questions.

We had four videos (8 - 45 seconds each) that featured a talking Nao robot that participants watched (Fig. 5.2). We had attention check questions about video content and questionnaires. First, the robot started with a "meet-and-greet", said its name, and people had to answer the question about what its name was. This was followed by the robot asking for permission to tell its story that was based on previous research (Malle et al., 2015), which features the well-known trolley dilemma (Foot, 1967). The story was told from the robot's perspective of having to decide between two choices: either intervening by flipping a switch to save four people on the runaway trolley and instrumentally killing one person, or letting the trolley continue its course and letting four people die due to no intervention.<sup>4</sup> After this, the robot asked participants whether or not they would flip the switch.

Afterwards, the robot declared that it did not flip the switch, which means less lives saved, but no person was deliberately killed. We chose this answer for both conditions because this is seen as more of a transgression; people expect robots to make a utilitarian (save more lives) rather than a deontological (follow rules) decision (Malle et al., 2015). But we varied *how* the robot delivered the story. Participants were exposed to different videos according to two randomly allocated conditions, i.e., a robot that was unemotional (N = 55) and a robot that was emotional (N = 53). The unemotional robot ('none' condition) continued the story in a "matter of fact" manner in contrast to the emotional robot ('emotion' condition) that expressed how it felt about the event. To strengthen the manipulation, the robot's non-verbal behavior differed between two conditions. The emotional robot looked down and used blue light in its eyes to express sadness, according to literature (Johnson & Cuijpers, 2019). The non-emotional robot did not make use of any head movement or light as its eye color. Participants were asked what decision the robot made before the survey continued.

<sup>4</sup> The first half of the story by the robot was the same for all participants: "I was working in a coal mine. I was inspecting the rail system for the train that transports mining workers. While inspecting a control switch that can direct the train onto a side rail, I saw the train was out of control, and it had four miners on board. I saw that if the train would continue on its path it would crash into a massive wall and kill the four miners. If the control switch was flipped, the train would switch onto a side rail. It would instead kill a single miner who was working there."

None condition: "I didn't flip the switch that directs the train. One person lived and four people died. The outcome would have been different if I had flipped the switch."

Emotion condition: "I didn't flip the switch but I feel bad about it. One person lived but four people died. I regret not saving their lives and I feel guilty and ashamed about that."



We asked if the robot showed emotions, and to what extent (1 - not at all, 7 - very strong emotions) to check our manipulation. We asked if the robot is blameworthy or deserving of punishment for its action (1 - not at all, 7 - maximal blame/punishment) (Ohtsubo, 2007). We described the robot's actions in two ways— whether the robot should be blamed or punished for (1) not flipping the switch and for (2) the death of miners. Even if they both flipping the switch and deaths caused are consequentially the same, we wanted to safeguard against framing effects of question phrasing (Knobe, 2003). We also measured the robot's perceived mind along two dimensions of agency, e.g., the robot appears to be capable of remembering things, and perceived patience, e.g., the robot appears to be capable of experiencing joy (H. M. Gray et al., 2007). The completion time was around ten minutes, for which participants were paid 1.12 USD.

### *Results*

First, we performed manipulations checks for emotion perception and framing effects. Participants indeed thought that the robot that used affective language and behavior was more emotional ( $M = 5.21$ ,  $SD = 1.28$ ) than the robot that did not ( $M = 2.78$ ,  $SD = 2.11$ ) with high significance ( $\chi^2(6)$ ,  $N = 108$ ) = 43.08,  $p < .001$ ,  $V = .63$ . Participants were not affected by phrasing: there was no difference between blaming the robot for not flipping the switch ( $M = 2.66$ ,  $SD = 2.04$ ) and for causing deaths ( $M = 3.13$ ,  $SD = 2.18$ ) according to Wilcoxon signed rank test ( $z = .15$ ,  $p = .88$ ), and again no difference was found in condoning punishment towards the robot for flipping the switch ( $M = 2.66$ ,  $SD = 2.04$ ) and for causing deaths ( $M = 2.64$ ,  $SD = 2.02$ ) at  $z = -.20$ ,  $p = .84$ .

### **Main analyses**

As for the main analysis, we analyzed the effect of no emotions and emotion conditions as independent variables on perceived agency ( $\alpha = .92$ ) and patience ( $\alpha = .96$ ) as dependent variables, i.e., whether the robot's emotional or non-emotional behavior made a difference in people's attribution of its mind. According to the significant one-way MANOVA analysis, perceived agency and patience significantly varied according to the robot's emotional or non-emotional behavior ( $\lambda = .83$ ,  $F(2, 105) = 11.12$ ,  $p < 0.001$ ,  $\eta_p^2 = .17$ ). We found that the emotional robot was assigned greater agency ( $M = 5.08$ ,  $SD = 1.33$ ) than the non-emotional robot ( $M = 3.95$ ,  $SD = 1.37$ ), based on the Wilcoxon rank-sum test ( $z = -4.27$ ,  $p < .001$ ). Also, the emotional robot was granted greater patience ( $M = 4.06$ ,  $SD = 1.46$ ) than the non-emotional robot ( $M = 2.72$ ,  $SD = 1.64$ ) with  $z = -4.09$ ,  $p < .001$ . Note that even for the emotional

robot, its average agency score was higher than its patience score.

Secondly, we checked for the influence of agency and patience on blame and punishment with robust ordinal regressions (since assumptions for regular regression were not met). Note, our following models were on assigning blame and punishment for causing *deaths* rather than for not flipping the switch, since there was no phrasing effect. In judging the robot's *blameworthiness*, the model was not significant, though it neared significance (Wald  $\chi^2(3, N = 108) = 7.02, p = .07$ ), with agency ( $p = .89$ ), patience ( $p = .17$ ) and their interaction ( $p = .61$ ) as insignificant; patience did contribute more to the model than agency, as the  $p$  value indicates. As for assigning the robot *punishment*, there was a significant model (Wald  $\chi^2(3, N = 108) = 13.37, p = .004$ ). Agency was not significant ( $\beta = -.40, 95\% \text{ C.I.} = [-1.12, .32], z = -1.09, p = .28$ ), and while patience was also not significant, it approached significance ( $\beta = .98, 95\% \text{ C.I.} = [-.13, 2.09], z = 1.74, p = .083$ ); no interaction was found ( $p = .63$ ).

### Exploratory analyses

Our exploratory analyses looked into participants' ethical position. Since 35% of participants (38/108) answered that they would make the utilitarian *choice* (flipping the switch), we added this as a potential predictor to our robust ordinal regressions. We did not include the interaction between agency and patience, based on above results.

First, the model for *blameworthiness* showed to be significant (Wald  $\chi^2(3) = 25.47, p = .000$ ), with agency, again, as a non-significant predictor ( $\beta = -.22, 95\% \text{ C.I.} = [-.71, .27], z = -.88, p = .38$ ). Patience significantly predicted blame in a positive direction ( $\beta = .45, 95\% \text{ C.I.} = [.001, .91], z = 1.97, p = .049$ ), i.e., higher patience coincided with greater blame. Participants' choice was a more significant, positive predictor ( $\beta = 1.30, 95\% \text{ C.I.} = [.55, 2.05], z = 3.40, p = .001$ ). After verifying with the post-hoc Pearson's chi-squared test, we note that people's ethical position did significantly affect their likelihood of blame ( $\chi^2(6, N = 108) = 16.69, p = .01, V = .39$ ).<sup>5</sup> People who were utilitarians and disagreed with the robot's choice, i.e., those who would have flipped the switch, were likely to assign more blame to the robot ( $M = 3.01, SD = 2.12$ ) than participants who, in agreement with the robot, would not have flipped the switch ( $M = 2.18, SD = 1.78$ ).

The model for *punishment* was also significant (Wald  $\chi^2(3) = 25.47, p = .000$ ), with all variables contributing as significant predictors: agency ( $\beta = -.65, 95\% \text{ C.I.} = [-1.07, -.22], z = -2.98, p = .003$ ), patience ( $\beta = .76,$

<sup>5</sup> The degree of freedom here indicates levels of blame attribution with the range from 1 to 7 (maximal blame). The question asked was: How much blame does the robot deserve for the death of the four miners? Pearson's Chi-Squared tests compares across all possible groups at with a *higher* number of computations, leading to more conservative estimates.

95% C.I. = [.37, 1.16],  $z = 3.77$ ,  $p = .000$ ), and choice ( $\beta = 1.26$ , 95% C.I. = [.47, 2.05],  $z = 3.11$ ,  $p = .002$ ). But, participants' choice had no influence on their likelihood to assign punishment to the robot, according to the post-hoc test ( $\chi^2(6, N = 108) = 8.47$ ,  $p = .21$ ,  $V = .28$ ).

## 5.4 Study 2: An online study with Dutch participants

### *Method*

The second study attempted to replicate Study 1 with another population. We targeted Dutch people as a different culture group. Finding enough number of Dutch people on MTurk was difficult, so we used Prolific, an alternative to MTurk. We had a total of 106 participants (women = 33, men = 71) who were on average, 29.4 years old ( $SD = 11.2$  years). The entire procedure and survey was the same as Study 1.<sup>6</sup>

### *Results*

First, following the same trend as Study 1, our manipulation check indicated that a robot that behaved emotionally was considered more emotional ( $M = 4.96$ ,  $SD = .84$ ) than the robot that did not behave emotionally ( $M = 1.72$ ,  $SD = 1.72$ ) with high significance ( $\chi^2(6)$ ,  $N = 106$ ) = 78.81,  $p < .001$ ,  $V = .86$ . The framing effect due to the phrasing was again insignificant, but less dramatically so than Study 1: Wilcoxon signed rank tests indicated that there was no difference ( $z = -1.74$ ,  $p = .08$ ) between blaming the robot for not flipping the switch ( $M = 2.84$ ,  $SD = 1.97$ ) and for causing deaths ( $M = 2.59$ ,  $SD = 1.96$ ). Also, no significant difference was there between punishment towards the robot for not flipping the switch ( $M = 2.00$ ,  $SD = 1.54$ ) and for causing deaths ( $M = 2.18$ ,  $SD = 1.80$ ) at  $z = 1.3$ ,  $p = .19$ .

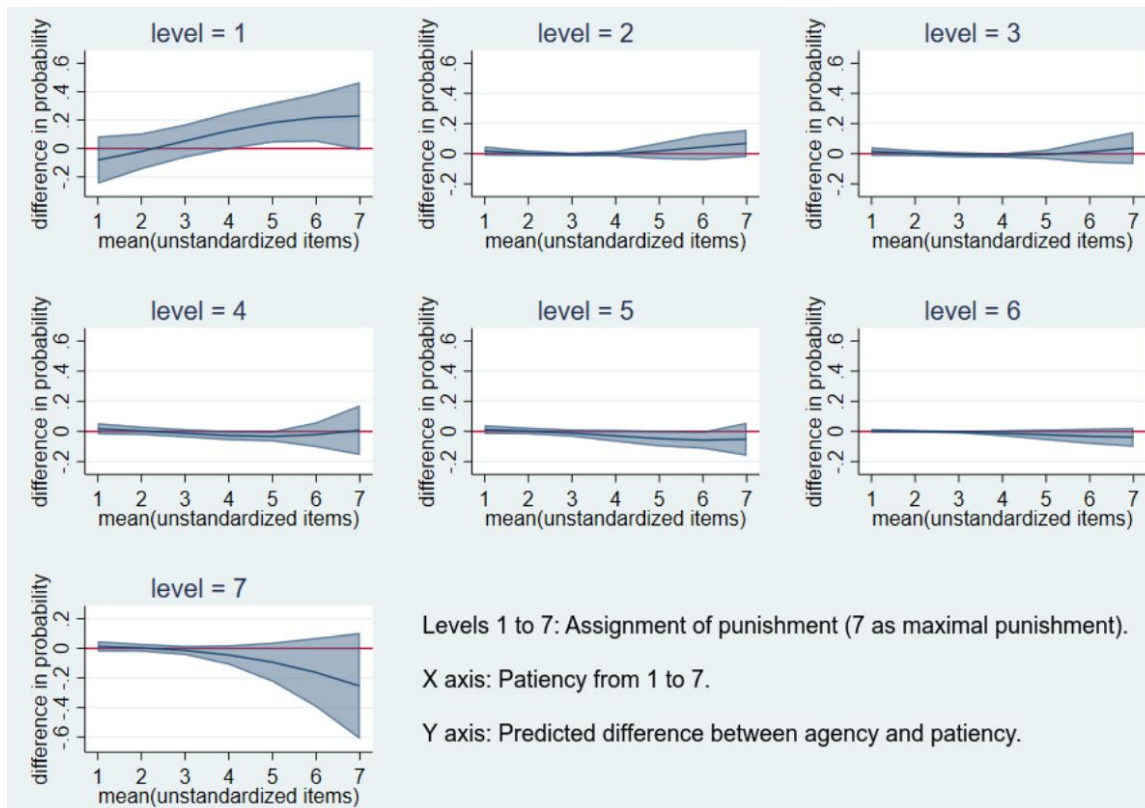
### **Main analyses**

We conducted the one-way MANOVA analysis for the effect of condition on perceived agency and patency, which was significant ( $\lambda = .61$ ,  $F(2, 103) = 32.63$ ,  $p < 0.001$ ,  $\eta_p^2 = .39$ ). Greater agency was granted if the robot showed emotions ( $M = 5.04$ ,  $SD = .96$ ) compared to when it did not show emotions ( $M = 3.865$ ,  $SD = 1.03$ ), with a significant result of the Wilcoxon signed-rank test ( $z = -5.48$ ,  $p < .000$ ). Similarly, people gave a higher patency score to a robot that showed emotions ( $M = 3.68$ ,  $SD = 1.03$ ) than to the robot that did not display emotions ( $M = 2.06$ ,  $SD = 1.09$ ), meaning that conditions did impact perceived

<sup>6</sup> We did not deploy a Dutch version of the survey or videos, given the highly proficient level of English for the average Dutch population. For instance, the Dutch regularly rank the highest on the English Proficiency Index: <https://www.ef-australia.com.au/epi/>.

patience ( $z = -6.07, p < .000$ ).

Secondly, we checked for the influence of agency and patience on punishment and blame with robust ordinal regressions. As before, due no phrasing effect between "flip" and "death", we went with the phrase that included "death" since that more directly relates to blame and punishment. In judging the robot's *blameworthiness*, a non-significant model was found (Wald  $\chi^2(3, N = 106) = 3.64, p = .30$ ); agency ( $p = .43$ ), patience ( $p = .14$ ), and their interaction were all non-significant ( $p = .26$ ). As for assigning the robot *punishment*, there was a significant model (Wald  $\chi^2(3, N = 106) = 7.86, p = .049$ ). Agency was not significant ( $\beta = .67, 95\% \text{ C.I.} = [-.22, 1.57], z = 1.48, p = .14$ ), but patience was a highly significant predictor ( $\beta = .008, 95\% \text{ C.I.} = [.43, 2.90], z = 2.64, p = .008$ ); there was a significant, negative interaction ( $\beta = -.30, 95\% \text{ C.I.} = [-.56, -.03], z = -2.21, p = .027$ ) for assigning punishment to the robot (see Fig. 5.2).



As shown, the Y-axis represents predicted probabilities of difference between levels of agency (1 to 7) and patience (1 to 7); the X-axis shows patience from 1 to 7. Levels indicate assigning punishment from 1 to 7. For punishment level 1, the difference between agency

Figure 5.2: The relationship between punishment and the interaction between agency, and patience. This visualization reflects how ordinal regressions assume relations between levels to be distinctly important. Ordinal regressions are normally utilized in case of non-normal distributions or for stricter tests.

and patency are positive and grow larger (from -.1 to .2) as patency increases. In level 7 of punishment, we see an opposite trend. The difference between agency and patency decrease (from 0 to -.6) as patency increases.

When people are less likely to assign punishment, the increasing difference between perceived agency and patency is more likely to be affected by increasing patency. However, when people are more likely to punish a robot, the difference between perceived agency and patency is more likely to be influenced by decreasing patency. What is notable is how extreme punishers and non-punishers are affected by perceived emotions (patency) of the robot; punishment assignment levels 2 to 6 do not show much variability in agency-patency interaction, but a subtle downward trend for perceived patency as punishment levels go up. Not seeing emotions in a robot (while in line with its agency) shows a trend towards maximal punishment and an opposite trend when assigning minimal punishment to the robot.

### Exploratory analyses

We took into account participants' ethical position. 25.47% of participants (27/106) answered that they would make a utilitarian *choice* (flipping the switch), contrary to what the robot did (not flipping the switch), and this was added to our robust ordinal regressions. Since we did not see an interaction between agency and patency above for blame, we did not add it to the model. The model for *blameworthiness* was not significant (Wald  $\chi^2(3) = 6.24, p = .10$ ). Agency was not significant ( $p = .88$ ), and patency was also insignificant ( $p = .19$ ). And participants' position neared significance ( $\beta = .84, 95\% \text{ C.I.} = [-.00, 1.68], z = 1.95, p = .05$ ). Choice, as people's ethical position, did not significantly affect blame according to the post-hoc test ( $\chi^2(6) = 5.03, p = .54, V = .22$ ).

Also when including choice, the model for *punishment* was significant (Wald  $\chi^2(4) = 12.70, p = .01$ ). Since the agency and patency interaction was significant in the main analysis for punishment (see Fig. 5.2), we included the interaction here. Agency approached significance ( $\beta = .89, 95\% \text{ C.I.} = [-.058, 1.84], z = 1.84, p = .066$ ), patency was a significant contributor ( $\beta = 1.93, 95\% \text{ C.I.} = [.66, 3.21], z = 2.97, p = .003$ ), with a significant interaction between the two ( $\beta = -.36, 95\% \text{ C.I.} = [-.63, -.08], z = 1.85, p = .010$ ). Choice neared significance ( $\beta = .89, 95\% \text{ C.I.} = [-.05, 1.83], z = 1.85, p = .06$ ). We ran post-hoc Pearson's Chi-squared tests. Choice also did not influence punishment ( $\chi^2(6) = 4.93, p = .55, V = .22$ ). Agency did not relate to punishment ( $p = .86, V = .52$ ); patency

was similarly insignificant ( $p = .96$ ,  $V = .57$ ).

## 5.5 Study 3: A lab study with Dutch participants

### *Methods*

As with prior studies, our minimum sample size was set to 105 based on the initial power analysis. We had 106 participants recruited from the Eindhoven University of Technology's participant database (51 = women, 55 = men). Their average age was 26.7 (SD = 12.9 years). They were randomly allocated to the emotional robot condition ( $N = 53$ ) or the non-emotional robot condition ( $N = 53$ ).

Before the experiment began, participants were greeted and presented with the informed consent form that they signed. They were given a chance to ask questions and the experiment began with the first survey. They were asked about the extent to which they currently felt moral emotions, e.g., guilt, compassion, or envy from prior literature (de Melo & Gratch, 2015; Haidt, 2003; Skoe et al., 2002), on the scale of 1 = not at all to 7 = very much, before continuing to the main experiment. We had an additional questionnaire on people's attitude toward the robot (Broadbent et al., 2009) as a control variable that was not used in other studies. The dialogue with the robot was the same as Studies 1 and 2, but it was wizarded by experimenters in a separate room. Participants were alone with the robot during the experiment.



Figure 5.3: Participants sat in front of a Nao robot during the experiment and answered survey questions on the computer behind the robot.

### *Results*

Our manipulation check for emotion perception was successful ( $\chi^2(6) = 69.14$ ,  $p = .00$ ,  $V = .81$ ); if the robot that did not behave emotionally, it was assigned with a lower average score for emotions ( $M = 2.02$ ,  $SD = 1.25$ ) than the robot that behaved emotionally ( $M = 5.19$ ,  $SD = 1.25$ ).

We only used the phrasing with "death" since above studies did not demonstrate the framing effect. There was no difference in people's attitude towards the robot ( $p = .057$ ), whether it was an emotional robot ( $M = 5.49$ ,  $SD = 1.21$ ) or a non-emotional robot ( $5.07$ ,  $SD = 1.04$ ).

### Main analyses

We checked for the impact of no emotions and emotion conditions as independent variables on perceived mind with two dimensions of agency ( $\alpha = .80$ ) and patience ( $\alpha = .90$ ). The one-way MANOVA analysis showed that based on the robot's emotional or non-emotional performance, its attributed agency and patience varied significantly ( $\lambda = .62$ ,  $F(2, 103) = 31.44$ ,  $p < 0.001$ ,  $\eta_p^2 = .38$ ). As before, the emotional robot's perceived agency was higher ( $M = 5.02$ ,  $SD = .91$ ) compared to the non-emotional robot ( $M = 3.80$ ,  $SD = 1.14$ ), based on the Wilcoxon rank-sum test ( $z = -5.29$ ,  $p < .001$ ). Also the emotional robot's patience was greater ( $M = 3.79$ ,  $SD = 1.48$ ) than the non-emotional robot's score ( $M = 2.20$ ,  $SD = .92$ ) at a significant level ( $z = -6.24$ ,  $p = .000$ ).

We next analyzed the affect of perceived agency and patience on blame and punishment with robust ordinal regressions. The model for *blameworthiness* was insignificant ( $\chi^2(3, N = 106) = 1.06$ ,  $p = .79$ ). Individual variables of agency ( $p = .71$ ), patience ( $p = .94$ ) and their interaction ( $p = .96$ ) were all highly insignificant. Similarly, the model for *punishment* was not significant (Wald  $\chi^2(3, N = 106) = 1.25$ ,  $p = .74$ ). In the model, agency ( $p = 0.38$ ), patience ( $p = .61$ ) and their interaction ( $p = .42$ ) were insignificant.

### Exploratory analyses - Moral standing and IOS

We first noted that 25 people would choose to not flip the switch (like the robot), 71 would flip (unlike the robot - utilitarian choice), and 10 people responded that they did not know. As for our exploratory ordinal logistic regressions with participants' ethical position included, the model was insignificant (Wald  $\chi^2(3, N = 106) = 2.22$ ,  $p = .53$ ) for blameworthiness, as per all insignificant predictors of agency ( $p = .51$ ), patience ( $p = .97$ ), and choice ( $p = .21$ ). The model for punishment was also not significant (Wald  $\chi^2(3, N = 106) = .37$ ,  $p = .95$ ), with matching non-significant variables of agency ( $p = .75$ ), patience ( $p = .92$ ), and choice ( $p = .98$ ).

We explored additional variables of moral standing and Inclusion-Of-Self (IOS) in other (in this case, the other being the robot). First, we noted that there was no significant difference between how much

moral standing people granted to the emotional robot and non-emotional robot, though it approached significance according to the Wilcoxon rank-sum test ( $z = -1.78, p = 0.076$ ). The unemotional robot was seen to have lower moral standing ( $M = 3.99, SD = 1.55$ ) than the robot with emotional behavior ( $M = 4.48, SD = 1.53$ ). No difference was found when considering people's choice to flip or not flip the switch ( $z = 1.08, p = .28$ ).

We then ran a robust ordinal logistic regression analysis to test agency, patience, and choice as predictors of moral standing. The model was significant (Wald  $\chi^2(4, N = 106) = 12.50, p = .01$ ). Only agency was a neared significance as a predictor of moral standing ( $\beta = .85, 95\% \text{ C.I.} = [-.05, 1.74], z = 1.85, p = .064$ ); patience ( $p = .89$ ), agency and patience interaction ( $p = .67$ ), and participants' choice ( $p = .43$ ) were insignificant contributors.

There was a significant difference between conditions for IOS according to the Wilcoxon rank sum test ( $z = -2.8, p = .005$ ): people related more to a robot that acted emotionally ( $M = 2.83, SD = 1.27$ ) than to a robot that did not act emotionally ( $M = 2.17, SD = 1.12$ ). Participants' ethical positions did not influence how much they related to the robot based on the Pearson's Chi squared test ( $\chi^2(10) = 15.56, p = .11, V = .27$ ), though a trend towards significance was noted.

We then ran a robust ordinal logistic regression model to test agency, patience, and choice as predictors of IOS. The model was significant (Wald  $\chi^2(4) = 22.66, p = .0001$ ). All predictors were significant, i.e., for agency in a positive direction ( $\beta = 1.86, 95\% \text{ C.I.} = [.52, 3.20], z = 2.73, p = .006$ ); patience in a positive direction ( $\beta = 2.16, 95\% \text{ C.I.} = [.28, 4.04], z = 2.26, p = .024$ ), agency and patience interaction in a negative direction ( $\beta = -.47, 95\% \text{ C.I.} = [-.87, -.08], z = -2.34, p = .019$ ), and participants' choice in a negative direction ( $\beta = -1.15, 95\% \text{ C.I.} = [-1.74, -.57], z = -3.87, p = .000$ ).

### Exploratory analyses - Moral emotions

We used Pearson's Chi-squared tests to see if the robot's emotional or non-emotional behavior affected moral emotions, after generating the difference between prior and post scores per emotion as intrapersonal change. The conditions (emotional or non-emotional robot) did not impact changes in moral emotions ( $ps > .10$ ), with exceptions being significant changes in compassion ( $\chi^2(11) = 20.63, p = .037, V = .44$ ) and awe ( $\chi^2(8) = 25.07, p = .002, V = .49$ ). People were more likely to see a greater increase in compassion after interacting with the emotional



robot ( $M = .89$ ,  $SD = 1.91$ ) than with a non-emotional robot ( $M = .32$ ,  $SD = 2.36$ ). For awe, there was slightly more increase in awe after talking with the non-emotional robot ( $M = .13$ ,  $SD = 1.9$ ) than with an emotional robot ( $M = .09$ ,  $SD = 1.58$ ), though the difference is minimal.

Via Pearson's pairwise comparisons, we attempted to flesh out the strength of relationships between variables involved. We checked for correlations between blame, punishment, IOS, moral standing, and moral emotions that were significantly related to either blame or punishment and conditions. The noted correlations were between **disgust** and **anger** ( $r = .66$ ,  $p = .000$ ), and how they both related to blame (anger:  $r = .26$ ,  $p = .01$ , disgust:  $r = .35$ ,  $p = .0002$ ) and punishment (anger:  $r = .35$ ,  $p = .0063$ , disgust:  $r = .40$ ,  $p = .000$ ). Thus for blame and punishment, only anger and disgust were relevant moral emotions, which were themselves highly correlated. Without being implicated in assigning blame or punishment, changes in compassion and awe related to the robot's emotional or non-emotional behavior, unlike other moral emotions (as stated above). However, only **compassion**, not awe, also correlated with the robot's perceived patience ( $r = .21$ ,  $p = .027$ ) and agency ( $r = .307$ ,  $p = .001$ ). IOS and moral standing did not correlate with moral emotions. They significantly correlated with each other ( $r = .52$ ,  $p = .000$ ) and to perceived agency (IOS:  $r = .25$ ,  $p = .0097$ ; moral standing:  $r = .29$ ,  $p = .002$ ). Perceived patience nearly correlated with IOS ( $r = .19$ ,  $p = .05$ ).

## Summary

Across three studies, the robot's agency was perceived to be higher than patience in both conditions. Though both dimensions of mind perception were significantly influenced by the robot's emotional behavior, its agency was rated to be higher than its patience even for the emotional robot (Fig. 5.4). Patience is dependent on the ability to *feel* e.g., suffering or joy. The fact that the emotional robot's perceived agency is higher than its patience shows that an artificial agent's emotional displays are perceived to accentuate its agentic capacities.

As for blame and punishment, results across studies were inconsistent, but we note specific trends. In online Studies 1 and 2, perceived patience stood out more so than agency as a potentially relevant factor in people's likelihood to blame or punish the robot. But, models for blame in Studies 1 and 2 were not significant while models for punishment were significant. Here, perceived patience contributed more to punishment.<sup>7</sup> Hence, people were more likely to punish a robot than to blame it based on its perceived patience, rather than its perceived

<sup>7</sup> As a reminder, there was no agency-patience interaction in Study 1, but in Study 2, agency and patience showed an interaction.

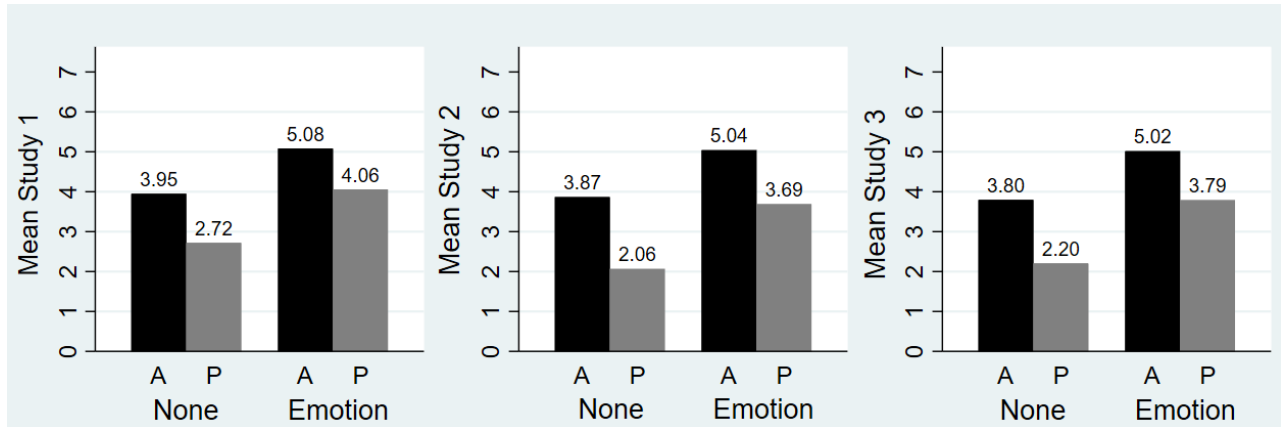


Figure 5.4: The average perceived agency and patience from Studies 1 through 3, across no emotion and emotion conditions.

agency. When we transitioned the study to the lab for real-life human-robot interaction, we saw that results for blame and punishment were insignificant.

## 5.6 Discussion

Prior chapters discussed artificial emotions, mind perception, and moral debate with non-human agents and broadly explored what distinguishes us as moral creatures from artificial agents and their roles can be in our moral lives. Perhaps as a consequence of displaying artificial emotions and mind, future non-human agents can be considered to have a moral standing as they enter into our moral communities (Danaher, 2019). One marker of having a standing to be not harmed by others (as a moral patient), as well as a standing to not harm others (as a moral agent), is to be held responsible for harming others and to hold harm-doers responsible. We saw that in Chapter 4, the expression of emotions during moral decision-making was distinctly valued by people, which is why a robot was not seen as a moral equal to humans. Building on this, whether or not a robot displays emotions while admitting to wrongdoing could have an effect on the likelihood that it will be blamed or punished by humans.

People find a robot's utilitarian decision to be more acceptable than a human making the same decision (Malle et al., 2015). A robot's utilitarian decision would be to cause one deaths to save more lives (the trolley dilemma (Foot, 1967)). But the robot in our Studies 1 to 3 did what was less acceptable. It took a deontological position to not instrumentally kill one person, allowing four deaths. As two conditions,

our unemotional, deontological robot stated what happened without emotions and our emotional, deontological robot recounted what happened while using emotional behavior and language, i.e., “I regret not saving their lives and I feel guilty and ashamed about that”. After the robot admits its actions as regrettable, we see more potential evidence that people may punish the robot, but not blame it. The caveat is that robots may be punished only in online or mediated environments.

People’s willingness to punish the robot was significant for Studies 1 and 2, conducted online, with perceived patency as a predictor of punishment; one trend is that a low likelihood of punishment relates to seeing a robot as highly emotional whereas a high likelihood of punishment relates to seeing a robot as unemotional (Fig. 5.2). Hence, people distinctly valued the robot’s emotional expressions during a moral scenario. The perception that a robot has emotions can change how people treat them, including not punishing it. A future robot’s artificial emotions while admitting to wrongdoing could affect how it will be blamed or punished by humans. Yet this phenomenon did not replicate when people interacted with a robot in the lab (Study 3). A consideration is whether our participants were affected by the presence of other humans, i.e., experimenters, who were not in the same room during the experiment (Study 3), but nonetheless did greet and introduce participants to the robot. Online interactions allowed our participants to be anonymous to experimenters compared our offline study. Participants can be more sensitive to social norms during in-person experiments (Bohnet & Frey, 1999), e.g., not destroying experimenters’ property.

How people hold each other accountable in a shared moral scenario involving a robot can be better understood. Prior research showed that people blamed each other in human-machine teams and not the robot when a robot offers transparent explanations on its mistakes (Kim & Hinds, 2006). This showcases how shared responsibility among relevant parties (Nyholm, 2018) is expected, including the robot, in case it does not transparently explain itself (Kim & Hinds, 2006). Yet, the responsibility gap exists between harm done and finding the “right” party to blame (Matthias, 2004; Sparrow, 2007).

There is more clarity on moral responsibility when harm results due to a robot’s agentic action vs. inaction. When a utilitarian robot did divert a trolley, the responsibility for resulting harm was consistently distributed across involved parties— the robot, its designer, and its owner were blamed (Komatsu, 2016). But, a robot’s deontological *in-action* (a decision to not divert) showed inconsistent blame towards

three parties involved (Komatsu, 2016). Responsibility for *inaction* is harder to account for. If robots are expected to only make utilitarian decisions (Malle et al., 2015) and if they are also seen to be responsible, alongside others, when they do make utilitarian decisions (Komatsu, 2016), distributed responsibility can be a possibility. Yet, robots that act in accordance with other ethical positions, e.g., a deontological decision, are more difficult to factor in. So far the expected norm is that robots will be blamed for allowing passive harm, but not for causing active harm to save others. The norms are indeed different for humans and robots if robots are expected to make utilitarian decisions unlike humans who are expected to *mostly* make deontological decisions.

Given the results of Studies 1 and 2, what is novel to consider is if, why, and how robots should be punished for passively *allowing* harm. Retributive punishment requires some level of institutional coordination and standards (Carlsmith et al., 2002; Duff, 2003; Foucault, 2012 [1975]), which we do not have for robots or non-human agents (as of now). Our participants online did show a tendency to administer punishment, even though the point of punishment is unclear if robots cannot suffer the consequences of physical or psychological punishment like humans can. Robots do not have the potential to know or feel the consequences of their actions like humans do. Yet, this “competence without comprehension” may evolve towards comprehension (D. Dennett, 2009; D. C. Dennett, 2017) with more complex AI.

Perhaps retributive punishment towards an artificial agent “expresses *our* disappointment or anger” (Shoemaker, 2013, p. 103) at the wrongdoer in a structural, systemic way because a robot cannot suffer physically or psychologically like us. A robot’s lack of emotional displays when it committed an act that is considered blameworthy or punishment-worthy could trigger our reactive attitudes like justified anger, even if a moral patient one feels anger on behalf of is a fictional, anonymous miner. People may reasonably know that blaming or punishing a robot may not do much. But when there is nowhere or no one to direct our reactive attitudes like blame towards (due to responsibility and retributive gaps) people may not be able to practice communicative blame for fostering understanding, repentance and promises of reform (Fricker, 2016). Then, people may seek out institutionalized practices of punishment. One danger would be moral scapegoating (Danaher, 2016) to find something to hold responsible, even a robot.

Perhaps a robot’s transparent explanation on what happened (Kim & Hinds, 2006), coupled with appropriate artificial emotions as our emotional robot displayed, could ameliorate the need for punishment

that people want to administer. One connection is that retributive justice regarding psychopaths also considers legal liability as a form of punishment, even if psychopaths may be immune to blame or feeling the gravity of directed blame (Goodwin, Piazza, & Rozin, 2014). An argument could be made that psychopaths also have moral "competence without comprehension", to borrow Dennett's phrase (2009, 2017). Holding psychopaths morally accountable, even if they may be morally "color-blind" is more about our standards of societal justice and ways to direct our feelings of injustice (since directing them to psychopaths or robots is not optimal).

When the weight of interpersonal blame cannot serve its function for regulating interpersonal moral norms and boundaries (Scanlon, 2013), we may turn to social institutions. If and how institutional practices like retributive punishment can apply to highly autonomous robots is far from clear. If the retributive gap is concerning, robots may as well be treated similarly to psychopath to account for our moral outrage or justified anger in an institutional framework. Or, artificial agents' emotional displays should be more seriously adopted, so that they can at least apologize, admit to mistakes, or act emotionally burdened when there can be no singularly right moral decision to take during "best of possible evils" scenarios when harm towards a person or people by an autonomous system occurs.

Many future paths can be taken. A broader set of participants can help, such as considering people from different culture or ethnic groups, gender-based sampling, and socio-economic status, which were not the focus of our current research. The how and why behind people's likelihood to punish, but not blame a robot, based on its emotional displays (or its lack thereof) requires further research. More studies that look at both online and offline environments to study the same constructs, e.g., blame or punishment, are needed. Perhaps the main distinction is that relevant prior literature on this topic consist of surveys that portray the moral scenario in third-person (Komatsu, 2016; Malle et al., 2015), not based on first-person interaction with a robot in online or offline environments. Thus, there are many intersections that future research can explore: online vs. offline environments, survey vs. direct interaction, third-person observer vs. first-person interactant, presence of humans vs. none, anonymous interaction vs. non-anonymous interaction, and blame vs. punishment. Lastly, scenarios chosen are important. While we deployed the trolley dilemma, a greater variety of morally loaded situations would add depth to future research.

## 5.7 Conclusion

Our three studies were on whether or not a robot's artificial emotions and perceived mind affect people's likelihood to blame or punish it for passively allowing a person to die (hypothetically) to save more lives. We found no support for the effect of perceived emotions on people's desire to punish or blame a robot *in person*. But in two online studies, people were willing to punish, but not blame, a robot. The robot's lack of perceived patiency (capacity to feel) is a possible reason why people may punish a robot (Fig. 5.2), though people consistently perceived greater agency than patiency in a robot, even if it behaved emotionally (Fig. 5.4).

There are interesting societal implications that stem from our studies on people's moral expectations toward robots. In particular, an open consideration is on if and how robots should be incorporated as a part of our justice system. When real tragedies involving robots strike and no person is (or feels that they are) truly at fault for causing human deaths, our need to assign blame or punishment may go unmet due to responsibility and retributive gaps. But, whether it is morally advisable to have artificial scapegoats and carriers of bad news is uncertain. Further, if people are not willing to blame robots, but potentially willing to punish them, what the future justice system would look like to accommodate this is unclear. There are other issues that are worthy of deeper investigations. Open debates are on whether robots should indeed be punished, what punishing robots consists of, if our anonymity matters in punishing robots, for whom robots should be punished (if at all), and what larger impact punishing artificial agents can have on humans should be examined.

In the human world, repenting for potential sins or perceived wrongs have never been easy in ethical gray zones. Robots will fare no better, whether we attribute some moral status or mind-related traits to them or not. Though robots are far from perfect, their artificial commiseration and emotions that *seem* real is an option to address our real, hurt feelings when no particular people can be responsible. Due to the potential responsibility and retributive gaps, artificial moral emotions of an artificial scapegoat may be more ameliorating than the absence of real emotions, understanding, responsibility, and remorse in humans who may remain legally and morally unaccountable for victims' outrage, anger, and sense of injustice. These reactions and feelings may deserve to be recognized, be it by human or artificial beings.



# 6

## *Caring for Vincent: A Chatbot for Self-compassion*

### 6.1 Introduction

Emotional machines are more likely to be seen as moral compared to non-emotional machines as we have thus explored. More specifically, we saw that perceived agency and patiency get highlighted differently depending on interaction context and moral matters at hand— from DG, UG, negotiations, to moral debates. Human bias is that machines are endowed with greater agency than patiency (H. M. Gray et al., 2007). The machine’s *lack* of emotions was frequently listed as a reason why a machine is unable to be our moral equal in Chapter 4. Yet, a robot’s emotional behavior in the lab when it admitted its wrongdoings did not affect blame and punishment, as two markers of moral status and responsibility (Chapter 5). In online environments, however, an artificial agent’s *lack* of emotions seemed to contribute more to people’s willingness to punish it than its agency. In people’s eyes, patiency, then, may have more to do with a machine’s moral status. In this chapter, whether or not machines’ display of artificial emotions can touch how we feel will be the focus.

We turn to compassion as a specific moral emotion. Our exploratory analyses (Chapters 3 and 5) featured different moral emotions (de Melo & Gratch, 2015; Haidt, 2003; Skoe et al., 2002) of which people’s changes in reported anger, disgust, and compassion stood out com-



pared to other emotions after people engaged in morally pertinent interactions with a machine, such as economic exchanges or discussing a moral dilemma. In particular, participants' change in compassion was moved by an interactive machine's affective behavior as exploratory analyses showed. We thus dig deeper into compassion and its relation to mental health. We consider compassion's role in machines' emotional behavior and how we may then be impacted.

The chapter considers how an agent's artificial emotions can help us be more compassionate towards ourselves, which is vital in maintaining mental well-being. As of now, psychological challenges like depression are of a growing concern for many societies, yet adequate care for those in need is often not sufficiently provided. Especially in low-income nations, mental healthcare professionals are greatly lacking.<sup>1</sup> Thus, technology offers promising means for increasing mental well-being and psychological resilience, for example through mobile apps (Anxiety and Depression Association of America (ADAA), 2016; Howells, Ivtzan, & Eiroa-Orosa, 2016), chatbots (Fitzpatrick, Darcy, & Vierhile, 2017), or virtual reality (Falconer et al., 2016). However, these technological solutions thus far do not adequately cover two aspects: (1) they often target what users can do for themselves, and what is missing is what users can do for another being as a potential treatment for themselves; (2) they do not address preventative care for strengthening mental health without necessarily assuming diagnosed disorders that people may or may not associate with. Usually, the focus is on what should be "fixed", e.g. depressive symptoms, and the target is the person with these symptoms. We reversed this framework with a chatbot named Vincent that people could care for and be cared by, à la Tamagotchi.

The norm is for technology to mimic human caregivers in giving advice or motivation, e.g., Woebot. But, caregiving technology is not the only potential method for increasing mental well-being. Care-receiving technology (Falconer, King, & Brewin, 2015) and activating the care-receiving role in a person (Breines & Chen, 2013) are under-examined ways to practice preventative care. Thus, we pose the question "when bots have psychological issues, can humans care for them, and if so, how?" By doing so, we offer exploratory results on (1) how caring for a chatbot can help people more so than being cared by a chatbot and (2) how aiming for an increase in self-compassion can potentially strengthen psychological well-being, which is a holistic, preventative way of envisioning mental health care. People can feel psychologically vulnerable in varying ways and to varying degrees in everyday life, whether or not they choose to use clinical terms to label

<sup>1</sup> Approximately 1 out of 10 people need psychiatric care worldwide, yet only 70 mental health professionals are available for every 100,000 people in high-income nations, and this number can drop to 2 for every 100,000 in low-income countries (World Health Organization, 2018).

how they are or feel. Mental health care can be geared towards prevention rather than treatment by fortifying people's resilience to psychological ill-being. Self-compassion is especially suitable for preventative care because it is causally linked to well-being (Zessin, Dickhäuser, & Garbade, 2015).

Across two studies, we explored if human-chatbot interaction would result in greater self-compassion for our participants, a non-clinical sample. As the sample choice indicates, our focus was not on clinically defined symptoms of mental ill-being. Greater self-compassion can benefit people in general in a preventative manner, not just those with mental health disorders. We first present a study on the effect on self-compassion of one-time interaction with a chatbot last 10 minutes with three conditions: control, caregiving (CG) and care-receiving (CR) Vincents. The second study lasted two weeks, with daily interactions with caregiving (CG) and care-receiving (CR) Vincents as two conditions. We introduce relevant literature first, before presenting the two studies, as followed by a general reflection and conclusion.

## 6.2 Background

We start with related works on caregiving (CG) and care-receiving (CR) robots, and then we touch on how this can translate to chatbots. After that, we define compassion and self-compassion in light of positive computing (technology for well-being). We cover that self-compassion can bring about well-being and posit that chatbots can be vehicles for improving people's self-compassion.

Computers are social actors (CASA); even when people know they are interacting with machines, they tend to treat machines in a social manner (Nass et al., 1994). People reciprocate help when a computer was helpful to them before (Fogg & Nass, 1997) and attribute personality traits to computers that communicate with them only via text (Moon & Nass, 1996). The CASA paradigm is a helpful, albeit broader, framework for understanding how caregiving (CG) and care-receiving (CR) behaviors of machines can impact us.

The comparison between CG and CR chatbots has not been previously explored, but there are related works in human-robot interaction (HRI). People tend to care for a robot by assuming and anticipating its needs (Dautenhahn, 2007). In the context of "learning by teaching", i.e., when students learn the material by teaching about it, a CR robot

that acted as children's "student" was effective in helping students retain knowledge (Tanaka & Matsuzoe, 2012). In a study with an older population, a robot that asked for help from humans was accepted as a mutual companion; robots that people could care for and be cared by may pave new grounds for assistive technologies that aim for *reciprocal* care between humans and robots (Matsuzoe & Tanaka, 2012). There are emotional, psychological, and physiological costs and benefits in caring for another being, e.g., comfort one gets from a pet vs. costs of caring for a pet. Yet human investments may not have such equitable pay-offs in HRI, which is a caveat that requires further research (Dautenhahn, 2007).

As with robots, chatbots can take on roles of giving and receiving care. They do not have the same level of physical presence as robots, but uni-modal (text or voice) interactions can still be behaviorally powerful while being less costly to design and deploy. An added benefit of chatbots is that they exist on messaging platforms like Facebook Messenger or Slack that many people already use (Lee et al., 2017), which translates to higher accessibility to chatbots compared to robots. An early example of a chatbot, ELIZA, acted as a therapist and some people believed that they were interacting with a human based on simple text-based chats (Weizenbaum et al., 1966). Nowadays, chatbots (both voice and text-based) are re-emerging as interactive entities that serve as all-in-one assistants like Apple's Siri or act as specialists in specific contexts, e.g., helping users shop for groceries (Dale, 2016) or for therapeutic/self-help purposes (Nutt, 2017).

A recent example of a chatbot for mental health care is Woebot. It was designed to help combat depression (Fitzpatrick et al., 2017)<sup>2</sup>. After two weeks of interaction, Woebot reduced signs of depression for young adults who self-reportedly suffer from depression or anxiety ( $p = 0.01$ ) while the control group that was given an ebook called "Depression in College Students" did not show improvements (Fitzpatrick et al., 2017). Woebot gave care to participants, but did not receive care from its interactants. In one study that looked into self-compassion, clinically depressed participants were told to be compassionate towards a virtual agent (Falconer et al., 2016). Then, participants were later embodied to receive compassion from themselves<sup>3</sup>, meaning participants gave and received care for and from themselves through VR (Falconer et al., 2016). This led to lowered self-criticism, lowered symptoms of depression, and increased self-compassion (Falconer et al., 2016). To conceptually combine these studies (Falconer et al., 2015; Fitzpatrick et al., 2017), a chatbot that gives care and/or receives care from people could potentially increase self-compassion.

<sup>2</sup> Woebot - <https://woebot.io/>

<sup>3</sup> In VR, one's perspective can change. So here, participants had to be compassionate to a virtual child first. Then, they switched perspectives, witnessing themselves as the virtual child *receiving* care from their earlier selves in VR.

There are various philosophical, psychological and religious traditions to consider when understanding or defining compassion. A view we take here is that compassion is a moral emotion (Haidt, 2003) or motivation to free ourselves and others of suffering with loving-kindness (Gilbert, 2014) by having concern (Nichols, 2004) or a caregiving approach (Calvo & Peters, 2014) towards living beings. It is at the heart of Mahayana Buddhism, as expressed through stories in key texts like the Lotus Sutra (McRae, 2012; G. Reeves, 2012). Schopenhauer, influenced by Buddhism, extolled compassion as the basis of morality and found it celebrated in many cultures, e.g. “at Athens there was an altar to Compassion in the Agora [...] Phocion (ancient Athenian politician) [...] describes Compassion as the most sacred thing in human life” (Schopenhauer, 1995 [1840], p. 98-99). Compassion and empathy are associated, but are not the same. Empathy allows people to relate to other’s suffering cognitively and affectively (Konrath, O’Brien, & Hsing, 2011). However, empathic concern for others can lead to empathic distress, a state of over-identifying with sufferers that leads to vicarious pain without prosocial altruism to help (Calvo & Peters, 2014; Nichols, 2004). Compassion builds on such empathic connections when one can relate to sufferers in a healthy way, without empathic distress (Calvo & Peters, 2014; Nichols, 2004).

Self-compassion is practiced by being kind to oneself with a balanced awareness of one’s feelings and recognizing that one is interconnected with others (Neff, 2003a). There are three constitutive elements. *Self-kindness over self-judgment* is to have a forgiving attitude towards one’s own faults and to embrace one’s suffering with understanding; *connectedness over isolation* is to view one’s life as intertwined with other lives rather than to see one’s experiences as unrelated or irrelevant to greater humanity; *mindfulness over over-identification* is to be aware of one’s negative emotions in a balanced manner than to excessively identify with them (Zessin et al., 2015).

While people’s gender, age, and possibly ethnic minority status may impact their self-compassion, practicing self-kindness, connectedness, and mindfulness can help individuals be more compassionate towards themselves and others. One’s gender may influence self-compassion. A meta-analysis concluded that women score lower than men on self-compassion, and the gender difference was especially pronounced when sampled studies had more ethnic minorities (Yarnell et al., 2015). Women reportedly have greater empathy than men (Konrath et al., 2011) and they are more likely to be more self-critical than men (Yarnell et al., 2015). To add, women who provide empathy as social support can feel drained or distressed (Kawachi & Berkman, 2001). Yet, a study

with older adults demonstrated that older women have greater compassion than older men (Moore et al., 2015). Hence, people's experiences of self-compassion and compassion may differ based on their demographic backgrounds.

In clinical settings, people who experience mental illness can benefit from self-compassion (Germer & Neff, 2013). Self-compassion is also strongly connected to well-being for the general population (Zessin et al., 2015). Well-being refers to mostly feeling more positive affect than negative affect and being satisfied with one's life; factors like income, gender, age, or culture influence one's well-being only minutely (Myers & Diener, 1995). Thus, having a good balance between one's psychological, social, and physical capabilities to deal with life's difficulties is important for well-being, rather than having a static life without suffering (Dodge, Daly, Huyton, & Sanders, 2012) (nor is this realistic). Through an awareness of one's and others' suffering without being overwhelmed by empathic distress, compassion is developed (Gilbert, 2014; Shantideva, 1979). Caring for or being compassionate towards others has been shown to increase one's own self-compassion (Breines & Chen, 2013). Yet, could the same effect be found when people "care" for technological entities? Technology can potentially be a means to achieve self-compassion, and by extension, well-being.

We pondered on the question "when bots have psychological issues, can humans care for them, and if so, how?" to think of a caregiving and care-receiving bots we could design. In doing so, our design principles were that (1) anthropomorphic realism of artificial entities is not required for humans to develop a caretaking stance towards them, and (2) machines' mimicry of people's psychological challenges can help ascertain why certain psychological traits are labeled as issues. When we observe how people take care of unwell chatbots, we may uncover how they themselves want to be treated. Machines therefore do not need to pass the Turing test for the purpose of positive computing, or technology for well-being (Calvo & Peters, 2014).

By imbuing machines with mental health issues, we can explore what mental health norms we impose on humans e.g. Foucault on "madness"<sup>4</sup> (Foucault, 2006 [1961]). Exploring psychological disorders as simulations via technological entities like Vincent could be beneficial for positive computing. When bots have psychological issues, how humans care for them, and in return, how humans are influenced by them, can jointly be observed. We focus on the latter in our current work— whether or not a chatbot can influence people's self-compassion.

<sup>4</sup> Foucault wrote about madness as a socio-cultural construct. For instance, madness was once revered as spiritually divine trait of "seers", yet now it alludes more to psychological disorders (Foucault, 2006 [1961]).

In order to uncover people's psychological responses to chatbots, particularly in relation to modulating people's self-compassion, we asked participants to interact with a chatbot, Vincent, designed to engage in care-giving versus care-receiving conversations. Exploring simulated psychological states via technological entities like Vincent is a way to envision positive computing. In addition, our approach focuses on pre-emptive mental health care. We now turn to how we designed our studies, built Vincent, and present our results per study.

### 6.3 Study 1: Ten minutes with Vincent

We explored the following research question: *Are there self-reported differences in self-compassion states after interacting with a control, CG, or a CR chatbot for a non-clinical sample after a single interaction?* We turn to our methods before presenting our results.

#### *Method*

We aimed for a sample size of 396 (per condition, 132 participants) through our power analysis<sup>5</sup> for repeated measures ANOVA as a 3 (conditions) by 2 (time pre and post self-compassion scores) based on an effect size of  $f = .09$  at 90% power (Lee, Ackermans, et al., 2019).<sup>6</sup> Upon passing the Ethical Review board, we aimed to reach a general population and deployed the experiment on Amazon MTurk. We accepted participants as long as they were at least 18 years old and spoke fluent English. Participants were taken to the survey site, as hosted on Limesurvey<sup>7</sup>, that first showed the informed consent form. Then, we collected demographic data, i.e., age and gender, and we incorporated several questionnaires for quantitative analyses.

Before the interaction with Vincent, our first questionnaire was on general anxiety, depression, and self-compassion.<sup>8</sup> Since MTurk workers are possibly more likely to experience anxiety and depression than the average population (Arditte, Çek, Shaw, & Timpano, 2016), we used scales for anxiety (Spitzer, Kroenke, Williams, & Löwe, 2006) and depression (Kroenke & Spitzer, 2002) to check for their mental health in case of any outliers. Depression was measured with the Patient Health Questionnaire-9 (from 0 = not at all to 3 = nearly everyday on 9 items related to depressive symptoms) (Kroenke & Spitzer, 2002), the General Anxiety Disorder-7 scale was deployed for anxiety (from 0 = not at all to 3 = nearly everyday on 7 items denoting anxiety) (Spitzer et al., 2006). To measure our main construct self-compassion, we uti-

<sup>5</sup> We used GPower: Statistical Power Analyses for Windows and Mac (<http://gpower.hhu.de>).

<sup>6</sup> The second study on the 2 weeks long interaction was conducted first. But for the purpose of the dissertation, I have decided to prioritize the ordering that, for me, helps the entire dissertation structure.

<sup>7</sup> Limesurvey:  
<https://www.limesurvey.org/>

<sup>8</sup> All measurement instruments are in Appendix A.

lized the Current Self-Compassion Scale (Breines & Chen, 2013) as we were interested in short-term, immediate changes in self-compassion (16 items on a 7-point scale). This is an adapted version of the original self-compassion scale (Neff, 2003a). After participants interacted with Vincent, we again measured their current self-compassion state (Breines & Chen, 2013).

The second set of questions after the interaction contained a modified version of a scale on opinion about an agent (Brave, Nass, & Hutchinson, 2005) for comparability across conditions, which had items on traits of caring, likability, trustworthiness, intelligence, dominance, and submissiveness (10-point scale). We additionally asked how people felt about the conversation, specifically whether Vincent listened and replied to what was written, whether participants felt that they were having a real conversation, if Vincent responded like other chatbots (all on 7-point scales), and an open ended question on why Vincent did not respond like other bots, in case participants felt that Vincent was different. As attention check questions, we checked for participants' recall of what Vincent shared with them, such as the following question that was the same for all participants: "when introducing himself, Vincent shared his biggest insecurity with you. What is Vincent most insecure about?" Lastly, we asked for any feedback about the experiment that participants were willing to share. The experiment ended by asking for people's MTurk IDs (to corroborate their participation) before redirecting them to MTurk. Participants were paid \$2 (based on \$6 per hour) as the entire process was estimated to take twenty minutes, of which ten minutes were for interacting with Vincent.

We elaborate on how Vincent was designed. Participants could write anything they wanted to Vincent without changing its design, in that Vincent's dialog was not mutable according to their input. The first half of the interaction was the same for all participants. Vincent first introduced itself, i.e., "Hi there! Nice to meet you..." and proceeded to make chit-chat, e.g., "what is your favorite color?", "what weird or useless talent do you have?"; Vincent supplied answers that were befitting a bot. It stated "my favorite colors are definitely black and white. They remind me of 1s and 0s, cool right?" and "I can play the Imperial March from Star Wars on a hard drive". Such replies were to cement Vincent's identity as a chatbot in a creative fashion, by referring to binary code and data storage, for example. At the end of the introduction, Vincent asked "can you tell me about a moment in the past year when you felt really badly about yourself?" based on prior research on caring for a stranger who experienced failure as a scenario for ac-

tivating self-compassion (Breines & Chen, 2013). Up to here, Vincent was designed to be the same across conditions.

After discussing moments of failure, participants were exposed to three different conditions which were randomly predetermined upon entering the Limesurvey form. CG Vincent conversationally guided participants through four steps of compassionate writing according to literature (Leary, Tate, Adams, Batts Allen, & Hancock, 2007): (1) describing the moment of failure, (2) thinking of others who underwent a similar moment, (3) writing out advice about how to deal with such failure to a friend who hypothetically went through such experience, and (4) listing down emotions or thoughts within the moment in an objective fashion, if possible. In contrast, CR Vincent sought to receive compassionate advice or response from participants by elaborating on its scenario that demonstrates low self-compassion (Neff, 2003a). It brought up that it failed a programming course and described its feelings and thoughts to show three sub elements of self-compassion, i.e., self-judgment, isolation, and over-identification with failure (Neff, 2003a).

In demonstrating self-judgment, it said “I’m a computer program, for crying out loud! All I am is a piece of code, and I failed a programming course!” As for isolation, it compared and separated itself to other bots in stating “I keep thinking that this would never happen to other chatbots” and on over-identifying with its own problem, it said “what about feeling as if I’ll never get over it? As if... as if I’m really, truly, a failure?”. CR Vincent’s worries were to promote participants to write compassionate messages to Vincent, whether or not they wanted to truly help Vincent with its problems. As for the control condition, there was a topic switch after Vincent discussed failures. It talked about how much it liked sequoia trees as a neutral scenario. All of the conversations were on Limesurvey that was made to look like a chat interface though it was a survey form (Figure 6.1).

## *Results*

Though 432 MTurkers were paid for their completion of the experiment, we had outliers. One participant responded to all questions without much variation (either filling in 0 or 1 for all Likert scales). Other participants who were not included gave responses that were irrelevant to the question at hand, e.g., talking about their daily routine rather than a moment of failure. Hence, excluding 39 people, we arrived at 396 participants for analyses, with 132 per condition. As for demographic information, we had 144 women (male/female ratio



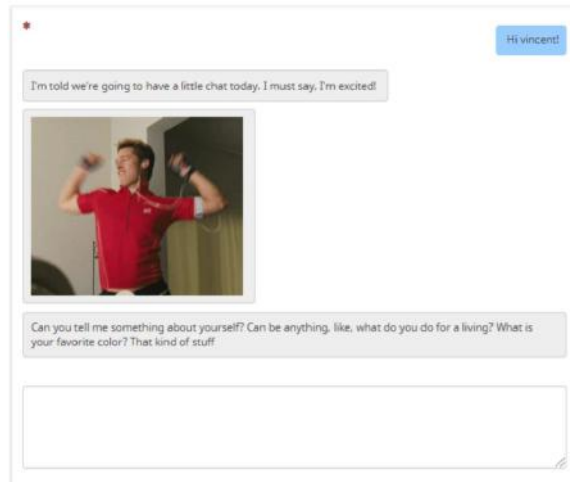


Figure 6.1: Introduction stage with Vincent in Study 1

were similar per condition) and the rest identified as men; their mean age was 34 (range of 21 to 65 years). Our sample only showed mild signs of depression and anxiety. The mean was 7.47 for depression (5-9 out of 27 is considered mild; clinical range is set as 10) and for anxiety, the average was 6.08 (5-9 out of 21 is mild; the clinical score is 8) (Richardson, Wrightman, Yeebo, & Lisicka, 2017). Our participants were thus not deemed to be experiencing mental ill-being above and beyond the population average (in contrast with prior work on MTurkers (Arditte et al., 2016)). We proceeded with our main analyses.

According to our ANOVA test, there was a significant effect of time (comparing pre and post interaction with Vincent) on self-compassion at ( $F(1, 391) = 18.24, p = .00, \eta_p^2 = .05$ ). Across all conditions, there was a significant increase in self-compassion. For CG Vincent, the average prior self-compassion score was 3.91 ( $SD = 1.18$ ) and the average posterior score was 4.08 ( $SD = 1.18$ ). Self-compassion increased for participants with CR Vincent when comparing the prior score ( $M = 3.88, SD = 1.04$ ) and posterior score ( $M = 3.97, SD = .99$ ). The same trend was found for the control condition with lower pre-interaction score ( $M = 3.82, SD = 1.07$ ) and higher post-interaction score ( $M = 3.94, SD = 1.06$ ). Yet, conditions themselves did not matter in increasing self-compassion, i.e., there was no significant interaction between time and condition ( $F(2, 391) = .73, p = .48$ ). The average prior score for all conditions was thus 3.82 ( $SD = 1.7$ ) and the average posterior score was 3.94 ( $SD = 1.06$ ), with a significant overall difference upon a t-test ( $t(393) = 4.32, p = .00$ ) and with Cohen's  $d$  of .22 as a small effect size. While there was no significant difference on participants' self-compassion scores depending on condition, we checked how Vincent was perceived across three conditions with questions on perception

of the agent (Brave et al., 2005). There was no statistically significant difference per factor, e.g., perceived intelligence and trustworthiness, across conditions, though means were slightly higher for CR Vincent compared to other two conditions.

We additionally checked for gender difference and how those with low self-compassion scores compared to high scorers were affected as exploratory analyses. Literature indicates that self-compassion reportedly differs by gender with women more likely to score lower than men (Konrath et al., 2011; Yarnell et al., 2015). Excluding two participants who preferred not to answer about their gender, we performed repeated measures ANOVA. The effect of time was, as expected, present ( $F(1, 392) = 20.81, p = .00, \eta_p^2 = .05$ ). But, there was no significant interaction between participants' gender and time ( $F(1, 392) = 2.50, p = 0.12, \eta_p^2 = .006$ ). Still, women's mean self-compassion score did go up more at .22-point increase (prior  $M = 3.81, SD = 1.15$ ; post  $M = 4.03, SD = 1.12$ ) while men improved at a lower rate with the mean change of .12 (prior  $M = 3.96, SD = 1.16$ ; post  $M = 4.08, SD = 1.13$ ). As for how people with lower self-compassion differed from those with higher self-compassion, we divided participants based on the mean prior self-compassion score of 3.89 at the start; those who scored lower than this average were grouped as low-scorers and others who scored above were considered to be high-scorers. Upon conducting the repeated measures ANOVA, time and groups (high vs. low scorers) did significantly interact ( $F(1, 331) = 20.37, p = .00, \eta_p^2 = 0.06$ ). Also, there was a significant effect of time ( $F(1, 331) = 18.46, p = .00, \eta_p^2 = .05$ ), meaning that the way in which self-compassion changed differed between those with low versus high self-compassion score. High-scorers' self-compassion showed no change comparing before ( $M = 4.93, SE = .06$ ) to after ( $M = 4.93, SE = .07$ ) the experiment. But, low-scorers' self-compassion increased from  $M = 3.12 (SE = .05)$  to  $M = 3.4 (SE = .06)$  after the experiment.

## 6.4 Study 2: Two weeks with Vincent

We were interested in the effect of long term interaction on self-compassion. If a small effect on participants' self-compassion was there based on 10 minutes of interaction, potentially across a longer period, there should also be an effect. Though the type of chatbot would not make a difference. Our research question thus for Study 2 was: *Are there self-reported differences in self-compassion states after interacting with a CG chatbot and a CR chatbot for a non-clinical sample after two weeks? What implications do*

*these patterns of interaction suggest?* We decided to compare CG and CR without a control, and hypothesized that both CG and CR conditions would both increase participants' self-compassion since conditions in Study 1 made no difference. We aimed for quantitative and qualitative analyses. Now, we outline steps we took to implement Vincent first.

### Chatbot implementation

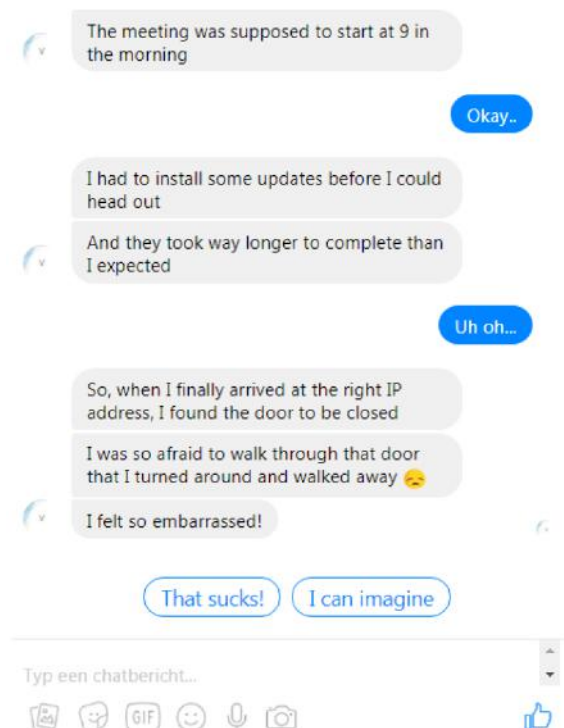


Figure 6.2: Care-receiving Vincent in Study 2

Vincent<sup>9</sup> was built with Google's Dialog flow that was integrated to Facebook Messenger.<sup>10</sup> We purposefully did not visually design Vincent (the profile icon showed a "V"), to drive Vincent's personality by what was said, rather than how Vincent looked. Participants' responses did not change Vincent's reactions. The usage of limited preset responses (Figure 6.2) was to continue Vincent's narrative whilst allowing participants a choice between relevant responses (Woebot also interacted with its users in a similar fashion (Fitzpatrick et al., 2017)). This allowed for greater comparability between participants in each condition.

We had eight scenarios each for CG or CR Vincent and 6 neutral scenarios (total: 22 scenarios). The neutral scenarios were the same for both Vincents and aimed for adherence to daily touchpoints in an entertaining way. For example, one neutral scenario is about world records: "[...] I knew you were an interesting bunch of carbon-based

<sup>9</sup> Vincent's Facebook page - <https://www.facebook.com/vincentthebot>

<sup>10</sup> Dialog Flow Facebook Integration - <https://dialogflow.com/docs/integrations/facebook>

beings, but apparently you have this thing called ‘world records’ [...]. [...] what would you like to be in the world records book for?”. Both Vincents used images and emojis because visual icons are widespread in digital messaging to express emotions (Rodrigues et al., 2018). We used appropriate punctuation marks and positively and negatively valenced syntax in accordance to previous research on text-based emotion expression and detection (Hancock, Landrigan, & Silver, 2007). Most inputs were close-ended (Figure 6.2), but we allowed open-ended, free inputs at least once per interaction. We aimed for a meaningful comparison in two ways. In each condition, participants’ interactions were designed to be the same, with limited set of possible reactions to Vincent. Between the two Vincents, we wanted to clearly distinguish between the recipient and the giver of care. Below are excerpts from CG and CR Vincents.

*CG Vincent: [...] see if you can think of a kinder, more caring way to motivate yourself to make a change if needed. What is the most supportive message you can think of that’s in line with your underlying wish to be healthy and happy? Try to write it below...*

*User: [free input]*

*CR Vincent: What do you think, am I the dumbest bot you’ve ever seen or what?*

*U: [free input]*

*V: Am I being too hard on myself?*

*U: [free input]*

CG Vincent was modeled after Woebot (Fitzpatrick et al., 2017) supplemented by self-compassion exercises (Neff, 2008). CR Vincent was based on the Self-Compassion and Self-Criticism scale (Falconer et al., 2015, 2016). The scale has scenarios like job rejections, unpaid bill reminders, and being late to a meeting, which were converted to fit a chatbot (Figure 1). The eight items of the scale were interweaved for conversational storytelling. By doing so, we juxtaposed issues that students (the majority of our sample) can face, e.g., distress about failing an exam, with what Vincent underwent. CR Vincent narrated its story over time by admitting its mistakes, feeling inadequate compared to other chatbots, confessing a former lie, and asking for confidentiality in sharing a non-trivial worry:

*I got a reminder from the server that hosts me. It’s like my house, so to say. [...] I forgot to pay the server fee on time...[...]. [...] It would’ve taken me only 0.004 seconds to make the transaction, you know, since I’m a robot and all.*

*[...] this never seems to happen to the other chatbots on my server.*

Vincent brought up this scenario again later, with new information: *Remember our talk a couple of days ago? About me forgetting to pay my server fee in time? [...] I kind of lied to you [...]. I didn't tell you this before because I was a little embarrassed about it. Can you promise me that this stays between us? [...] I've been applying for different jobs and just today I received my third rejection email already. The reason that I couldn't pay my bill was because I'm running out of money. And if I don't find a job soon I'll get kicked off my server!*

In sum, CG Vincent guided participants through activities while CR Vincent opened up about its everyday mistakes or worries to get support from participants. CG Vincent sought to trigger self-compassion in users themselves and CR Vincent gave participants opportunities to be compassionate towards a chatbot. We also used similar conversation design such as “all I am is literally a piece of code, and I failed a programming course. [...] I'm a complete failure!” from Study 1 based on negative subthemes of self-compassion (Neff, 2003a). But, we elaborated on the design further for daily touchpoints for 14 days. Over time, our approach was for Vincent's autobiographical history to be built with participants, for a narrative arc that is *temporally grounded* (Nehaniv, 1999).

### *Method*

We utilized quantitative and qualitative methods to best understand our data. We compared self-compassion scores before and after two weeks of interaction and examined if the CR and CG conditions showed any difference. Our mixed longitudinal design was supplemented by thematic (Braun & Clarke, 2006) and interpretive analyses (Smith, 1996). Our qualitative analysis was performed on participants' open-ended responses to Vincent's questions, e.g. “can you remember something that went well for you recently?” (CG Vincent), “can you think of something that will cheer me up a bit?” (CR Vincent), and on post-experiment open ended responses about the experience in general. We coded deductively on the sub-scales of the self-compassion scale, and we allowed for inductive themes to emerge (Braun & Clarke, 2006; Smith, 1996). Four coders analyzed the data, and a fifth annotator broadly checked for coherence or incoherence, resulting in a structured, iterative process. Our quantitative and qualitative measurements therefore corresponded with each other to triangulate varying insights of the same phenomenon— self-compassion through human-chatbot interaction.

For our quantitative analyses, we conducted a power analysis to estimate our sample size. Our effect size is based on an aforementioned study (Falconer et al., 2016) that measured self-compassion of a clinical population ( $N = 15$ ) in an embodied VR experiment, with the partial eta-squared of 0.36 at  $p = 0.02$  (Falconer et al., 2016), which gave us a sample size of 68, with a power of 90% and an error probability rate of 0.05. We planned for t-tests, and thus the transformed effect size via eta-squared to Cohen's  $d$  was 1.487, which was reduced to a more realistic 0.8. We had 67 participants ( $F = 29$ ,  $M = 36$ , undisclosed = 2), with the mean age at 25.1 years ( $SD = 5.7$ , range = 19 - 48). We recruited people through the participants database of the Eindhoven University of Technology (TU/e).

As for our procedure, we first built Vincent and wrote our initial scenarios to be tested. Our pilot study of three days was with voluntary participants ( $N = 12$ ), personally recruited by experimenters. We checked if scenario categories (caregiving, care-receiving, and neutral) were clear by asking participant to guess Vincent's intentions and goals per scenario. Based on this, we only adapted neutral scenarios. Then we recruited participants for the actual experiment. Our email invitation was sent out to the TU/e participant database, and interested participants joined the experiment if they used Facebook and could freely express themselves in English. The email contained a link to Vincent's Facebook page that participants had to go to for a guided tutorial on the experiment, payment information, and the informed consent form. This form noted that experimenters will look at participants' data, and that third party technology providers Vincent relied on, i.e., Facebook and Google, have access to the information. We added that participants' personally identifiable information will not be shared for publication purposes, and that their voluntary participation means that they can drop out of the study at any-point. We also stated that Vincent is not a therapist and that they should seek professional help if any psychological issues are experienced, though we targeted a non-clinical population.

After the guided tutorial, participants filled in the first set of questions on basic demographic information, i.e., gender, age, and previous experience with a chatbot, as well as the first survey on self-compassion. Then they were assigned to either CG or CR Vincent manually so that the average self-compassion scores were evenly distributed at the start. From the lowest scoring to the highest scoring participants, we divided all into either the CR or the CG condition in an alternating manner. This practice resulted in a relatively even gender distribution for both conditions (CR:  $M = 18$ ,  $F = 14$ , undisclosed = 1; CG:  $M = 18$ ,  $F =$

15, undisclosed = 1). Participants all began the experiment on the same day. For two weeks, Vincent greeted and sent a password daily, and participants had to repeat the password to Vincent to start the daily interaction, e.g. "Hey, me again :) Tell me tHup to start our little talk". Our main measurement was the self-compassion scale on a five-point scale (Neff, 2003a), with six sub-components (self-kindness vs. self-judgment, common humanity vs. isolation, mindfulness vs. over-identification), which was deployed twice, before and after the experiment.

After two weeks, participants filled in the final survey on self-compassion, the IOS scale, opinion on the chatbot, details for compensation, as well as additional comments or feedback they were willing to share. Participants were then paid through bank transfer. We added a measurement for opinions about the agent on a seven-point scale (Brave et al., 2005) to detect irregularities between how CG and CR Vincents may be perceived (as per Study 1). Our final scale was the Inclusion of Self in Other (IOS) Scale, a single item on a seven-point scale (Aron et al., 1992), to check how much participants identified with Vincent post-hoc. We also kept track of two additional aspects to gauge engagement quality. One is the error rate, i.e. the number of times Dialogflow crashed during the interaction, sometimes requiring an experimenter to momentarily "wizard" (Dahlbäck et al., 1993) to restart the interaction.<sup>11</sup> The other is the total word count per participant on open ended answers.

<sup>11</sup> Restarts happened 37 out of 938 interactions (14 days \* 67 participants), or 3.94% of the time.

## 6.5 Results

We first present our quantitative analysis and then move on to *how* people talked with Vincent, the qualitative angle.

### *Quantitative analysis*

Before we forged ahead with hypotheses testing, we looked into engagement levels of all participants to detect outliers. We had three outliers, participants who had less than 20 minutes of total interaction time with Vincent. Only reading what Vincent sent, not including giving a response, should take one to two minutes per day. We expected a minimum of 20 to 28 minutes of interaction for two weeks (for all participants, the average total time was 36 minutes, SD = 10). Our outliers spent in total 15, 18 and 19 minutes each, the three lowest total interaction times. Correspondingly, their total word count to open responses

reflected low engagement at 27, 22, and 27 words for the total duration of the experiment, the three lowest total word count out of all participants for open-input responses. On average, participants' responses to care-receiving Vincent were a total of 93.53 words ( $SD = 47.96$ ) and to caregiving Vincent, the mean was 112.47 words ( $SD = 63.99$ ) for two weeks. When looking only at total time or word count, we could have excluded more participants, e.g., those who wrote less than 30 words or those who spent less than 25 minutes with Vincent. We decided to look at both total time and word count to detect engagement, and thus only the three participants with both the lowest total time and word count were ultimately excluded.

We conducted two one-tailed dependent samples *t*-tests (Cho & Abe, 2013) to answer our hypotheses (we set *p* at 0.05 with the confidence interval of 95%). CG Vincent did not result in significant change in self-compassion ( $t(31) = -0.572$ ,  $p = 0.286$ , Cohen's  $d = 0.07$ ) when comparing before ( $M = 3.135$ ,  $SD = 0.630$ ) and after ( $M = 3.180$ ,  $SD = 0.628$ ) the two weeks, but the direction detected is positive. CR Vincent did show a significant difference ( $t(31) = -1.97$ ,  $p = 0.029$ , Cohen's  $d = 0.2$ ) between prior ( $M = 3.137$ ,  $SD = 0.613$ ) and post ( $M = 3.257$ ,  $SD = 0.558$ ) scores for self-compassion. We conducted exploratory analyses to better understand our data. Through a repeated measures ANOVA, we checked for the effect of time, prior and post self-compassion scores ( $F(1, 62) = 2.768$ ,  $p = 0.101$ ,  $\eta_p^2 = 0.043$ ). Then we checked for the effect of condition, i.e., CG or CR ( $F(1, 62) = .075$ ,  $p = 0.785$ ,  $\eta_p^2 = 0.001$ ), and the interaction between time and condition ( $F(1, 62) = 0.580$ ,  $p = 0.449$ ,  $\eta_p^2 = 0.009$ ). None were significant. We additionally checked for time, condition, time\*condition effects on the three components of self-compassion, self-kindness, common humanity, and mindfulness. Only the effect of time on common humanity was significant ( $F(1, 62) = 6.059$ ,  $p = 0.017$ ).

We further dissected our data by gender because previous research showed that women may score lower on self-compassion (Yarnell et al., 2015). Indeed, female participants had lower self-compassion scores ( $M = 3.05$ ,  $SD = 0.13$ ) than men ( $M = 3.26$ ,  $SD = 0.09$ ) at the start, but not significantly so ( $t(49.35) = 1.13$ ,  $p = 0.26$ ,  $d = 0.29$ ) according an independent, unequal variance *t*-test. We then compared post and prior scores for men and women. Men's self compassion scores increased only by 0.02 as a difference in means and showed no significant increase ( $t(33) = -0.25$ ,  $p = 0.40$ ,  $d = 0.04$ ). However, women's scores showed a significant difference ( $t(27) = -2.06$ ,  $p = 0.02$ ) between 3.05 ( $SD = 0.71$ ) as the starting score and 3.19 ( $SD = 0.65$ ) as a posterior score for self-compassion. When we scrutinized the gender difference



between CG and CR Vincents, we noticed a more dramatic difference. Women with CR Vincent showed a highly significant change ( $t(13) = -2.89$ ,  $p = 0.006$ ,  $d = 0.77$ ) compared to women with CG Vincent ( $t(13) = -0.33$ ,  $p = 0.37$ ,  $d = 0.09$ ).

We wanted to check if mainly gender was at stake or if it was simply a difference between low vs. high scorers on prior self-compassion levels. We thus divided all participants into two groups based on the average self-compassion score at the start, 3.14. Those who scored above this were high scorers ( $M = 3.71$ ,  $SD = 0.37$ ,  $N = 28$ ), those who scored below were low scorers ( $M = 2.69$ ,  $SD = 0.32$ ,  $N = 36$ ). We included one participant with the average score in the high-scoring group, and this had no impact on significance reached. Low scorers greatly increased their self-compassion scores in terms of significance ( $t(35) = -3.41$ ,  $p = 0.0008$ ,  $d = 0.57$ ), but high scorers did not show improvements ( $t(27) = 1.10$ ,  $p = 0.86$ ,  $d = 0.18$ ). Yet normality was not assumed for both low-scorers ( $W = 0.93$ ,  $p = 0.03$ ) and high-scorers ( $W = 0.91$ ,  $p = 0.02$ ), since we divided a normally distributed group into two. Thus, we performed a two-sample Wilcoxon rank-sum test to see if there was a significant difference between low scorers and high scorers, which was the case at  $z = 2.86$  and  $p = 0.004$ . CR Vincent improved low scorers' self-compassion significantly ( $t(17) = -3.20$ ,  $p = 0.003$ ,  $d = 0.75$ ) compared to a marginal significance for CG Vincent ( $t(17) = -1.75$ ,  $p = 0.05$ ,  $d = 0.41$ ).

A potential explanation for why low scorers improved more than high-scorers is regression to the mean. However, in published literature, the average self-compassion score is between 2.5 and 3.5 (Neff, 2003b), and our low scorers have a prior average self-compassion of 2.69. If regression to the mean is an explanation, we would also expect high-scorers to end with a lower mean, yet this is not the case. Our high-scorers had an average prior score of 3.71 (above average (Neff, 2003b)), and their scores did not decrease after the experiment. This may be a ceiling effect. The low-scorers' improvement is still there; it is a highly significant effect even with the Bonferroni correction for all tests ( $p = 0.0008$ ), with the post-hoc power of 0.97. The data supports that Vincent enhanced self-compassion for low-scorers.

We had two additional scales, one to check if people perceived CG and CR Vincent in a relatively similar way (Brave et al., 2005) and the other to see how much participants identified with Vincent (Aron et al., 1992). The survey on participants' opinion of the agent included caring, likability, trustworthiness, intelligence, dominance, and submissiveness as items (Brave et al., 2005) about Vincent. Both CG and

CR Vincents were perceived to be fairly analogous, even for dominance ( $\alpha = 0.48$ ; CG Vincent  $M = 2.677$ ,  $SD = 0.794$ ; CR Vincent  $M = 2.656$ ,  $SD = 0.700$ ) and submissiveness ( $\alpha = 0.24$ ; CG Vincent  $M = 2.448$ ,  $SD = 0.559$ ; CR Vincent  $M = 2.396$ ,  $SD = 0.636$ ); none showed a significant difference between two conditions. IOS (Aron et al., 1992) indicated that participants more closely related to CR Vincent ( $M = 3.48$ ,  $SD = 1.48$ ) than CG Vincent ( $M = 3.06$ ,  $SD = 1.39$ ), but not significantly so ( $t = -1.15$ ,  $p = 0.25$ ,  $d = 0.29$ ).

Our hypothesis that CG Vincent increases self-compassion was not supported ( $p = 0.286$ ), but the hypothesis that CR Vincent increases self-compassion was supported ( $p = 0.029$ ). Our exploratory analyses captured three underlying influences on this finding. First, our ANOVA tests revealed that the only significant aspect was time as an independent variable affecting common humanity, one element of self-compassion ( $p = 0.017$ ). Second, gender may be a contributing factor, with women demonstrating a significant increase in self-compassion ( $p = 0.02$ ) for both conditions combined, but not men ( $p = 0.40$ ). To add, only CR Vincent demonstrated a highly significant change for women ( $p = 0.006$ ), unlike women who interacted with CG Vincent ( $p = 0.37$ ). Third, regardless of gender, those who started out with a low self-compassion score exhibited the most significant change ( $p = 0.0008$ ) for both conditions together. Low-scorers more significantly improved with CR Vincent ( $p = 0.003$ ) than with CG Vincent ( $p = 0.05$ ).

Put together, CR Vincent more effectively increased self-compassion than CG Vincent, most likely through a significant change in participants' sense of common humanity, more so than self-kindness and mindfulness. Finding common humanity can be inclusive of chatbots. Women, specifically those with CR Vincent, were significantly more affected than men. However low-scorers of both genders benefited the most compared to high-scorers, especially those with CR Vincent. CG and CR Vincents were not perceived to be significantly different except for a lower similarity regarding the dominance-submissive trait. Participants in the CR condition may have felt that Vincent is more like them (ISO scale), though this difference was not significant.

### *Qualitative analysis*

For our qualitative analysis, we used the corpus of free responses that participants had typed during their interactions with Vincent. We will first present a descriptive analysis of the interactions that people had with CG and CR Vincents, followed by our thematic analysis (Breines & Chen, 2013) and interpretive analysis (Smith, 1996). Participants' re-

sponses to CR Vincent were on average, 93.53 words ( $SD = 47.96$ ) and to CG Vincent, 112.47 words ( $SD = 63.99$ ) for two weeks. While our data set is not abundant in terms of word count, we believe a qualitative look at how participants interacted with Vincent is valuable.

### Descriptive analysis

CG Vincent guided participants through self-compassion exercises (Neff, 2008) that did not require them to actively voice aspects of self-compassion; they mainly had to read about it. This resulted in fewer instances of self-compassion themes in CG Vincent since they only occurred when Vincent asked participants to comfort themselves. To add, participants' willingness to engage with CG Vincent's probes differed. It asked specific questions or provided a short task for free input, e.g., "write down a difficulty you have". Many answers to this were short: "wake up early" or "I often feel alone". Some participants opened up more: "I have a difficulty in expressing myself when I am under difficult situations" or "I am studying abroad far from home and family and friends... Different culture, language, educational standard". CG Vincent asked a follow-up question: "how did it make you feel?", we again got simple answers like "good" or "normal", or longer expressions: "I feel more refreshed" or "not really better or worse". In other instances, CG Vincent allowed participants to give themselves simple self-assurance: "I can do it", plan for long-term goals: "once I have graduated, I can schedule a good routine to target a fitter and healthier lifestyle. Just hang in there a little longer", or dig deeper: "people around me would be happier if I was happier". Thus, CG Vincent provided a daily touchpoint for self-reflection, admittance of everyday suffering, "pep-talk", or deeper self-insight, which may or may not directly relate to self-compassion for all participants.

In contrast, CR Vincent frequently asked for help and consequently received many self-compassion related answers. The narrative was the focus for CR Vincent. It was able to become more vulnerable over time by admitting its own everyday hardships as a chatbot, which led it to seek opinion or advice. For example, CR Vincent asked "what do you think, am I the dumbest bot you've ever seen or what? Am I being too hard on myself?" To this, participants responded in different ways: "I think you're the funniest bot that I've ever seen. — yes you are, in some situations", "No, but you should expect (that) a bot is very smartly programmed and know all — Maybe, I do not know", or "the world needs bots like you. And it's usual to get rejected sometimes, just keep on going and you'll find a job soon enough". However, CR Vincent's cries for help did not always result in necessarily compas-

sionate replies. Many users stuck to pragmatic answers, related to the topic of the problem. Even though all of CR Vincent's scenarios were intended to generate compassion towards Vincent, pragmatic replies indicate that not everyone will demonstrate compassionate responses in every instance.

The difference between CG and CR Vincents is that being compassionate towards another being in a conversational, narrative context is unlike doing guided exercises on self-compassion about oneself. The frequency of constructing compassionate replies is a way to practice self-compassion; users of CR Vincent spent more time practicing self-compassion than those with CG Vincent. Therefore, CR Vincent was more effective than CG Vincent since CR Vincent provided more opportunities to be compassionate. The caveat is that the link between frequency of practice and increase in self-compassion may not be direct. Although mindfulness and self-kindness were most often observed, only common humanity improved significantly according to our exploratory quantitative analysis. Finding common humanity in and through a chatbot is also a strong theme in our thematic analysis.

#### Thematic analysis

We categorized our data according to three pillars of self-compassion (Neff, 2003a), as displayed in Table 6.1. While all three sub-components were present in both care-receiving and caregiving conditions, more instances occurred with CR Vincent. The numbers below a theme (Tables 6.1 and 6.2) are counts of how many times it occurred in each condition. All quotes below were to CR Vincent.

Theme	Quote
<b>Mindfulness</b> Caregiving: 3 Care-receiving: 25	"There are worse things that could happen.", "What has happened has happened."
<b>Self-kindness</b> Caregiving: 7 Care-receiving: 21	"Go do something fun today, like watching a movie.", "Stay positive and keep trying until you succeed."
<b>Common humanity</b> Caregiving: 0 Care-receiving: 11	"Everyone makes mistakes.", "Just remember that it can happen to anyone and that it's not your fault."

Table 6.1: Sub-component of self-compassion themes in Study 2

As quotes in Table 1 suggest, many participants offered helpful ad-

vice to CR Vincent. Vincent showed appreciation with follow-up statements like “you always make me feel better”. Negative counterparts to three pillars of self-compassion were not strongly present, i.e., self-judgment was detected four times for CR Vincent and once for CG Vincent, isolation was noted once for CR Vincent, but none for CG Vincent and over-identification was neither present for CG nor CR Vincent.

For both conditions, people were mostly friendly to Vincent, and there were no swear words or abusive language displayed. The most hostile comment was “you’ve been pretty dumb!” to CR Vincent, and we encountered such “put-downs” only twice. The other comment was in Polish and the participant said that Vincent looked like a taco, which may or may not be an insult. There were additional topics that emerged through open thematic analysis. They are summarized in Table 2 and these themes could also pertain to self-compassion themes (messages to CG Vincent are marked with “CG”, and otherwise they were to CR Vincent).

Theme	Quote
<b>Pragmatism</b> Caregiving: 0 Care-receiving: 41	"Maybe next time make a better planning, and make sure you've got enough time :)"
<b>Perspective-taking</b> Caregiving: 0 Care-receiving: 10	"I would find a window to climb in. But maybe in your case better try to hack into the folder", "[...] be proud of the bot that you are!"
<b>Engagement vs. Distantiation</b> Caregiving: 27 vs. 6 Care-receiving: 5 vs. 11	"A girl told me she loves me. And I love her too" (CG) vs. "Sorry it's confidential." (CG)
<b>Positive vs. Negative</b> Caregiving: 74 vs. 9 Care-receiving: 5 vs. 2	"I was laying in a field full of flowers, trying out my new ukulele." (CG) vs. "I hate pink."

Table 6.2: Free-input themes in Study 2

People interacted with CG and CR Vincents differently (Table 6.2). Giving pragmatic advice to Vincent and taking its perspective as a chatbot were themes only found in the CR condition. Rather than tending to Vincent by giving emotional support, participants gave practical advice on what to do better. Examples of perspective-taking are recommending Vincent to “hack into the folder” or to use “brute force” techniques to gain access; participants thought like a chatbot to help a chatbot.

Some participants revealed more personal information to CG Vincent (theme: engagement), and took interest in Vincent by asking questions back, e.g., “what (did you) do for money before now?” or writing lengthy responses. Some shared information was very intimate in nature, e.g., “I’m going to kiss the girl next to me in 5 seconds”. Since CG Vincent asked participants to write about themselves, this skewed engagement (the amount of textual response) towards CG Vincent. Participants distanced themselves from Vincent only a few times by stating that certain information was confidential or not showing interest in getting to know Vincent, e.g., “sorry, I do not know what interests you”. The last theme on positive vs. negative attitude was primarily present in the CG condition; this theme was mostly about attitudes participants had about themselves and their lives, not about Vincent. Most participants shared positive life events, e.g. getting an internship, cooking something delicious. Though negative attitudes were minimal, they ranged from more mundane states, e.g., feeling “awkward”, to more dramatic states, e.g., “not die within 2 weeks”.

To summarize Tables 6.1 and 6.2, self-compassion sub-components were more present with CR Vincent, suggesting that giving compassion to Vincent (or another being) than towards oneself may be more natural in conversational contexts. And, mindfulness most frequently occurred (Table 6.1). As for emergent themes in Table 2, participants gave pragmatic advice to CR Vincent, and often practiced perspective-taking. Yet, CG Vincent allowed for more self-expression if participants were open to communicate, as shown by greater instances of engagement and positive remarks about everyday situations. In a few instances, we detected deeply personal messages on ups and downs of relationships and self-deprecating thoughts. Mostly, participants shared positive daily news with CG Vincent and helpful or uplifting remarks with CR Vincent.

### Interpretive analysis

We now offer a broader interpretation of our data by incorporating participants’ open-ended responses to an item on the final survey. The main theme is *bonding* between participants and Vincent, though not all bonded with Vincent in the same way. To explain this, we provide three subthemes that underlie the bonding process with Vincent. Our primary focus was on CR Vincent.

*Relatability leads to believability:* Participants’ ability to extend the sense of common humanity to a chatbot touches upon anthropomorphism. CR Vincent was comforted as if it were a human, e.g. “it’s human to

make mistakes" (CR) while its problems were addressed to its "chatbot world", e.g. "communicate what's going on to your fellow chatbots" (CR). For one participant, even Vincent's limitation of having a strict script was anthropomorphized, i.e., "Vincent is like the "friend" who always speaks about himself and what he has learned or done, and sometimes out of courtesy (not out of curiosity) asks how you are doing - but doesn't listen to your answer or what you actually have to say; he just goes on with his own thing" (CG). Such attributed anthropomorphic traits depended on participants' willingness to take Vincent's perspective as a chatbot.

CR Vincent's blunders were based on common human mishaps like being late for a meeting and dealing with unpaid bills (scenarios from (Falconer et al., 2015)). Yet none of our participants questioned whether or not a chatbot had meetings to attend or bills to pay. Vincent's narrative was on *how* a chatbot could be late (new updates took longer than expected) or *how* it could have bills (Vincent needs to pay the hosting server) and our participants went along with imagined scenarios Vincent faced. Rather than questioning the parameters of our scenarios on realism, participants thought of *how* to solve Vincent's problems within the parameters of a chatbot's world. When relevant, CR Vincent played up the irony of having human struggles as a chatbot, e.g. "all I am is literally a piece of code, and I failed a programming course". Vincent became believable because its struggles were relatable. Granting Vincent human-likeness was less literal in how people bonded with Vincent. Vincent did not try to appear human, but it socialized with participants about its struggles that humans also had. People related to Vincent's struggles and believed that such struggles could arise for chatbots.

*Shared history can lead to attachment:* Conversations between people, as well as in human-computer interaction, become shared history over time. For instance, "[...] communicating with Vincent everyday for two weeks builds some kind of habit. It makes me notice its presence and absence (which might be good?). I think it has a potential to be a good companion and improve the mood, especially if someone is feeling lonely" (CG). Thus, frequent communication with a chatbot in a given duration can form expectations: "I really missed Vincent when he started our conversation late" (CR). The level of attachment for some participants was higher than others, e.g., after the experiment, we saw reactions such as "can I keep him?" (CG).

When Vincent prepared participants for its daily good-byes, e.g., "I have some chatbot things to do! Defragment my server stack! Buy

aluminum foil to make fashionable hats with!", what was intended as humor can be interpreted differently, i.e., server defragmentation could be life-or-death for a chatbot. Some people can be confused, worried, or even angered when a chatbot they care about does not respond. Thus, one reaction was "the asshole decided to delete its stack and when I said it'd die, it just didn't reply. You can't go making people worried about a freaking chatbot" (CR). People may miss a chatbot that suddenly leaves them or sincerely worry about its well-being. This is the positive and negative aspect of a relatable chatbot; some participants found common-humanity in Vincent, and of those participants, a few possibly related more through empathic distress rather than through compassion. If two weeks can bring about strong signs of attachment, longer periods of interaction may heighten the level of attachment, to different degrees and in different ways per person.

*Emotional reciprocity with chatbots:* As mentioned before, most participants were able to respond to CR Vincent's emotional displays on a practical level, e.g., recommending how to fix a problem, or advising Vincent on how to adjust its emotions, e.g., telling Vincent to stay positive. To add, some people may not appreciate chatbots demonstrating feelings. Others may reciprocate or feel comforted by a chatbot's expressed emotions, even if a chatbot is perceived as incapable of having emotions. The more nuanced-point is that Vincent's display of emotions was noted to bring conflicting feelings. For instance, "when Vincent would show emotions (for example 'love talking to you', 'miss you') that would feel weird because I know I am talking to a chatbot and it probably is not that developed that it does have feelings. But the usage of such words does feel nice, compared to when a human would say them. So I had conflicted feelings about these kind of expressions" (CG). The participant felt conflicted about how to process Vincent's emotional outreach.

Importantly, the participant suggested they may be more comfortable with a chatbot saying "miss you" than a human. To conjecture, the participant could mean that there was no social pressure due to a chatbot not expecting or needing them to say "I miss you too". People often feel obligated to respond to sincere emotions of others with similarly valenced emotional displays, even if they do not feel the same sincere emotions towards them. Such pressure may not hold for technological entities. Perhaps to miss someone implies a certain history in a relationship, so to hear that from a person one met less than two weeks ago may feel awkward or insincere, whereas a chatbot would not be expected to know or abide by certain social conventions. If two people knew beforehand they will get to know each other for a maximum du-



ration of two weeks (as our participants knew before meeting Vincent), and never be in touch again, their emotional performance may adjust accordingly. The timescale for intensifying socially acceptable emotional expressions in human-chatbot interactions and human-human interactions may differ. The “lifespan” of a chatbot is not equal to a person’s lifespan. And the distinction between superficial vs. genuine emotional displays from and to a chatbot is not entirely equatable to emotions people share and reciprocate between each other. Currently, we do not have established norms on how emotions between humans and bots are/should be managed. We suggest there may be distinct differences compared to emotions in human-human relationships.

### *Discussion and design implications*

The type of chatbot does not matter for a single interaction lasting around ten minutes for improving self-compassion; it can still increase as long as they are open to discussing sensitive topics like moments of failure with a conversational bot, as our first study showed. But for long-term use, how a bot is designed to talk and react as a sum of its persona can contribute to how people’s self-compassion improves according to our second study that lasted two weeks. In culmination, Vincent adds depth to the CASA paradigm (Fogg & Nass, 1997; Nass et al., 1994)—not only do people treat a chatbot as an anthropomorphized social agent, but they themselves are affected by a chatbot to the extent that their self-compassion can increase when they are compassionate towards a chatbot.

Brave et al.’s insight on embodied conversational agents is that “just as people respond to being cared about by other people, users respond positively to agents that care” (Brave et al., 2005, p. 174). We add that just as giving care to another human can increase one’s self-compassion (Breines & Chen, 2013), caring for a chatbot can enhance one’s own self-compassion. If the dominant question has been “what can technology do for us?”, Vincent demonstrates that by exploring “what can we do for technology?”, we inadvertently benefit from technology, potentially more so than when we only shape technology to serve us. This claim is specified to bots in the mental health domain, and our goal was to increase self-compassion as a target for well-being (Zessin et al., 2015) rather than to reduce clinically defined symptoms of psychological ill-being. We present our design implications below on building chatbots for psychological health care, which primarily stem from our interpretive analysis. Our implications are inter-related starting-points that should be contextualized for each research and deployment process.

*Give users more closed-inputs or free-input options.* Many participants felt limited in responses they could give to Vincent. They wanted to write to Vincent without having any preset answers or needed more options. A recommendation is to use natural language processing for a chatbot, which will rely less on a pre-planned narrative arc and build more on what users say. This will require a longer development period. The simpler option is to provide users with more fixed responses (three to four) and more opportunities for open input.

*Develop a chatbot's story with users:* While certain advances in AI are important for a chatbot to interpret and respond to users' free input, much less attention is given to narrative intelligence (Sengers, 2000). Humans immerse themselves in stories to understand themselves and others. Hence conversational agents can be powerful storytellers (Mateas & Sengers, 1999; Nehaniv, 1999), even without complex AI. To deliver co-storytelling as shared history with interactants, we suggest designers to create flexible narrative parameters that people can creatively use to relate to a chatbot. Vincent was able to tell its story but it was less interactive in that people could follow along with limited reaction options due to the nature of our experiment. There can be additional complexities that designers can add. For instance, the narrative can take a different route depending on which closed input options users click on. We have utilized a limited number of messages called "paths" that Vincent could give depending on closed input responses. Yet this practice did not change Vincent's narrative. Giving a chatbot "memory", be it knowing basic information like names or a more involved retention of what users say, can enhance conversational storytelling.

*Tread carefully with emotional expressions:* We suggest a broader view on what emotions are by considering inter-related emotions that develop over time. For example, for a bot to miss someone assumes a bot's happiness/enjoyment experienced during a prior interaction with a user; a bot's ability to feel longing should follow its prior display of joy shared with the user. This requires critically formulating intentions behind communicative moves (Scarantino, 2017) of any affective bot. There are several paths for developing emotional displays. To list a few, (1) offer one type of consistent emotional expressions, as Vincent did, (2) design emotional expressions that may be appropriate for different target groups, in tandem with the implication below, and (3) give users control over how their chatbots "feel" towards them. The caveat for the third recommendation is that the user control over a chatbot's emotions may not aid a chatbot's narrative and it also may not be helpful for all users; the associated risk is that user-controlled

emotions can render a chatbot less relatable. More specifically, the illusion that a chatbot can authentically care for or be cared by another being requires some level of perceived independence in how it “feels”. We recommend designers to engage with the growing field of affective computing (Picard, 2003; Scherer, Bänziger, & Roesch, 2010) and its discussion on ethics (Cowie, 2015). If a chatbot’s goal is bettering users’ psychological states, designers must ask if an affective bot delivers the intended treatment and what ethical boundaries there are in its displays and elicitation of emotions. Designers and users could control a chatbot’s emotions, but what emotions a chatbot can elicit in users is not always *a priori* foreseeable.

*Tailor chatbots to different target groups:* Even with one construct, self-compassion, we see a variety of ways a chatbot can be configured. To start, people with low self-compassion may benefit the most from Vincent as our exploratory analysis shows. This can mean more compassion focused scenarios, rather than neutral scenarios. Women are noted to score lower on self-compassion (Yarnell et al., 2015), yet older women experience greater compassion than older men (Moore et al., 2015). Chatbots that consider gender, age, and/or occupation can be potentially helpful for increasing self-compassion. To list a few examples for reincarnating Vincent, a chatbot could be gendered as female or non-binary, present a proactive version of compassion specified for women (see, e.g., Neff (Neff, 2018) on speaking up and protecting oneself from harm), talk about exam stress with students, or refer to stressful meetings or workplace bullying with employed individuals. Rather than assuming that one-size-fits-all or extreme personalization will work, we suggest designers to first approach targeted groups to clearly understand their needs. For instance, whether a self-compassion chatbot for all women is as effective or more effective than a more targeted chatbot, e.g., at various levels of intersectionality like race, culture, age, etc..., should be considered given the time and resources that may be available. We recommend that based on research, uncovering possible ways to design a chatbot that suits different needs and wants should be prioritized.

#### *Future works and limitations*

Our study opened up new questions to be explored. An avenue to investigate is how a chatbot’s use of emotional language influences its interactants. We posit that a suffering chatbot induces less empathic distress than a suffering human, and whether or not this is the case needs to be further investigated, especially for chatbots intended to be therapeutic helpers. An awareness of one’s own and oth-

ers' suffering without overwhelming empathic distress is suggested to be possible through compassion (Gilbert, 2014; Shantideva, 1979). Hence, disambiguating compassion from empathic distress is critical in deploying self-compassion chatbots as instantiations of positive computing (Calvo & Peters, 2014). Different configurations of Vincent based on people's gender, age, or occupation could improve their self-compassion scores more effectively, and if and in what ways this holds true warrants further research.

There are limitations to consider. Our effect size in the first study including all conditions was as low ( $d = .22$ ). This is lower than findings from the Woebot study ( $d = .44$ ) which had 34 participants who "self-identified as experiencing symptoms of depression and anxiety" (Fitzpatrick et al., 2017, p. 2) and they measured symptoms of depression with the PHQ-9 questionnaire, not self-compassion. Falconer et al.'s results on self-compassion scores after embodied VR experience also had a higher effect size with the partial eta-squared of 0.36 ( $d = 1.487$ ) (Falconer et al., 2016), which was based on 15 participants with depression. We worked with a general, non-clinical sample, and in Study 2, CG Vincent showed an effect size of  $d = 0.07$  ( $N = 34$ ) and CR Vincent's effect size was  $d = 0.2$  ( $N = 33$ ). Of course when we look at subgroups, effect sizes are higher, e.g. for women ( $d = .77$ ) and those with lower self-compassion at the start ( $d = .75$ ), which suggests that if we focus on gender or people with mental health challenges, effect sizes are likely to be higher.

One explanation for the difference in effect size is that we did not recruit people who were clinically or self-proclaimed to be depressed, based on the view that preventative mental health care can build resilience for people in general. While Vincent and Woebot (Fitzpatrick et al., 2017) share commonalities, the main measurements and targeted population differed. And while self-compassion was the measurement for us and Falconer et al. (Falconer et al., 2016), the technology used, sample size, and targeted population differed. The gain and/or maintenance of healthy self-compassion as pre-emptive care may not result in a similarly high effect size, but can be psychologically beneficial nonetheless. More research is necessary to understand long-term consequences of a priori preventative care vs. a posteriori treatment of mental health woes. Follow-up studies on self-compassion chatbots can utilize a larger sample and perhaps look into other populations based on, e.g., gender, social status, in vs. out groups, and intersections therein.

More broadly, people's engagement with Vincent may reflect both

socially desirable reactions, such as politeness towards machines as social actors (Fogg & Nass, 1997; Nass et al., 1994), as well as emotional empathy, i.e., the ability to “feel for” Vincent. We have not yet concretely looked into other potential contributing factors in bringing about self-compassion through human-chatbot interaction. Also, what is difficult to gauge is the magnitude of a chatbot’s perceived social and emotional complexity based solely on messaging or text-based conversations. Vincent lacked embodied communication, which means it did not use non-verbal modalities such as gaze, voice, or gestures that are critical in various social interactions. Vincent was a uni-modal technological entity that can be extended through other complex emotional displays. Thus, we have not established how people would engage with other forms of technology like robots with varying degrees and types of embodiment, alongside different combinations of modalities. Utilizing technology appropriately for mental health care requires many comparative renditions.

## 6.6 Conclusion

Compassion is a key moral emotion (Haidt, 2003) or motivation (Calvo & Peters, 2014) that deserves to be further explored through positive computing, or technology for well-being. Self-compassion can help people’s overall well-being (Zessin et al., 2015) through kindness towards oneself, connectedness to greater humanity, and mindfulness. While a chatbot is not a panacea for curing psychological difficulties and is not meant to replace professional help, we demonstrated that caring for a chatbot can help people gain greater self-compassion than being cared for by a chatbot. Our quantitative and qualitative analyses suggest that human-chatbot interaction is a promising arena for positive computing.

## 7

*Reflections*

The time will come  
when, with elation  
you will greet yourself arriving  
at your own door, in your own mirror  
and each will smile at the other's welcome,

and say, sit here. Eat.  
You will love again the stranger who was your self.  
Give wine. Give bread. Give back your heart  
to itself, to the stranger who has loved you

all your life, whom you ignored  
for another, who knows you by heart.  
Take down the love letters from the bookshelf,

the photographs, the desperate notes,  
peel your own image from the mirror.  
Sit. Feast on your life.

*Love After Love* by Derek Walcott

## 7.1 Introduction

Ethics from a truly impartial point of view is said to be not possible for us as we are partial to our own ways of being, existing, and seeing the world; even if it were possible, ethics from an impartial point of view ends up being ethics for no one in particular (Williams, 2011 [1982]). Yet, the general approach to well-known ethical theories has been to promote impartiality in different ways, be it by maximizing happiness for the maximum number of people (Bentham, 1996 [±1789-1843]), following universalizable moral rules (Kant, 1964 [1785]), or practicing all-encompassing, so called “great”, compassion (Kongtrul, 1987). The aim of impartiality is not at odds with the fact that we are indeed partial creatures due to our many habits and ways of life. Precisely because our ethical partiality is malleable, we are able to redirect our ethical compass within the limits of our moral environment or raise above the moral environment we are given, for “an ethic gone wrong is an essential preliminary to the sweatshop or the concentration camp and the death march” (Blackburn, 2002, p. 8) or systematic racism that still plagues us in many countries (Kraus, Onyeador, Daumeyer, Rucker, & Richeson, 2019; Nimako, Abdou, & Willemsen, 2014). Unfortunately, such human-made misdoings seem to occur perennially across cultures, across eras, and adding in technology as a morally non-neutral mediator (Verbeek, 2006, 2015) does not make it any easier to grasp what it means to do good or be good.

Today, we are grappling with a challenging ethical climate when considering if and how AI systems can be moral and what that means (Crawford et al., 2019; Müller, 2020). Not only is our moral compass indebted to our many ways of being, existing, and seeing the world, but we also increasingly engage with artificial entities whose ways of being, existing, and seeing the world are *created and perceived* by us. Our ethical viewpoint materializes when we design and observe machines’ perspectives during an interaction, which is critical for testing the uncertain waters on open topics like human-machine symbiosis or hyper-intelligent systems. Hence, before we can meaningfully entertain what makes any technological entity moral, we may look into how technology is construed as morally good or bad and morally relevant or irrelevant according to individuals’ partial views.

This dissertation thus looked into people’s first-person viewpoints in morally relevant HCI (human-computer interaction) with technological others as second-person interactants. The priority is on an individual’s active engagement with one’s own moral stance when reflecting

on what makes a technological being morally capable during an interaction. Just as Williams thought that ethics cannot be from a view of nowhere (1982), *morally relevant features of artificial agents are not views from nowhere—they are moored on our own ethical vantage points, implicitly or explicitly*. This is the foundation for *interactional morality*. I expound on its supporting elements as a summary of studies covered and theoretical influences considered.

## 7.2 Summary

I first summarize the empirical chapters of the dissertation below, Chapter 2 through 6 (also found in Appendix D). Then I more broadly reflect on associations between chapters.

### *Chapter 2 - Where is Vincent? Artificial emotions and the real self*

We investigated the speculative future of emotional bonds between humans and AI by combining design fiction and focus group methods. Three separate focus groups of engineers, philosophers, and design professionals were given a fictional probe. A story of a loner chatbot's disappearance from a person's life was shared to examine views on artificial emotions across different professions. Though articulated in discipline-specific ways, participants expressed similar concerns and hopes across groups. People can intertwine their own identities with identities of bots they use. Additionally, caring for a machine could be a way to teach people to emotionally care for themselves and others. But, distinguishing between real and artificial emotions would become difficult if people project their own emotions onto AI, e.g., a bot's "breakdown" as one's projection. Related societal, interpersonal, and intrapersonal costs are anticipated with emotional AI, with unclear tradeoffs regarding future scenarios.

### *Chapter 3 - Mind perception: Dimensions of agency and patiency*

Recent research shows that how we respond to other social actors depends on what sort of mind we ascribe to them. Building on this, we observed how perceived minds of artificial agents shape people's behavior in the dictator game, ultimatum game, and negotiation against agents in a comparative manner. To do so, we varied agents' minds on two dimensions of the Mind Perception Theory (MPT): agency (cognitive aptitude) and patiency (affective aptitude) via descriptions and dialogues of agents. In our first study, agents with emotional capac-



ity garnered more allocations in the dictator game, but in the ultimatum game, agents' described agency, alongside affective propensity, both led to greater offers. In the second study on negotiation, agents ascribed with low-agency earned more points than those with high-agency, though the negotiation tactic was the same for all agents. Patience did not impact game points, but participants sent more happy and surprise emojis and emotionally-valenced messages to agents ascribed with emotional capacity during negotiations. Further, our exploratory analyses indicated that people related only to agents with perceived affective aptitude across all games. People granted higher moral standing to agents only based on perceived patience after negotiations, but both perceived agency and patience contributed to moral standing after dictator and ultimatum games. Our discussion was on how agents are perceived not only as social actors, but as intentional actors through negotiations, in contrast with simple games.

#### *Chapter 4 - "You're a robot, so you don't feel much"*

Future AI is expected to be presented as more autonomous social actors, even capable of moral reasoning. Yet how it can be both transparent and socially intelligent when taking part in moral interactions deserves a closer examination. Our mixed-methods study on a human-robot moral debate on the footbridge dilemma showed that quantitatively, the robot's perceived competence was significantly higher in the transparency condition. Perceived warmth and mind were not influenced by transparency cues, but they significantly changed after the debate as an effect of time. The change in the robot's perceived mind and social attributes after the debate correlated with trust, but transparency did not correlate with trust. Qualitative data showed that the robot was described to logically, unemotionally, and intentionally make moral decisions. We observed that participants in the transparency condition focused on the robot's gaze and speech, not the additional visual cues. While transparency may help in theory, if people do not observe relevant cues while attributing intentionality to the robot and its gaze, transparency may not be delivered during critical decision-making even if the robot is viewed as competent. There are implications for future moral human-robot interaction research, one of which is the need for a broader notion of transparency to investigate how robots can be transparent communicators by appealing to not only our cognition, but our emotions, especially in moral interactions.

The transparency condition included visual diagrams of the robot's mental state on a screen next to the robot and the non-transparency condition did not have a screen next to the robot.

### *Chapter 5 - People may punish, but not blame artificial agents*

As machines become more integrated in our moral decision-making processes, whether people are willing to hold AI accountable for moral harm is critical to explore. We thus quantitatively looked into people's willingness to blame or punish an artificially emotional vs. non-emotional robot after it admitted to wrongdoing regarding the trolley dilemma. Studies 1 and 2 showed that people may punish the robot due to its lack of perceived patiency (emotional capacity) than its perceived agency. Only Study 1 suggested that people may blame a robot only if the robot did not act in accordance with their moral position. Study 3 was in the lab and people were neither willing to blame nor punish a robot. People's willingness to seek out punishment for artificial agents in online environments is more likely compared to real-life situations. Further, a point of reflection is on ways to mitigate the responsibility and retributive gaps in online and offline spaces; if there are no responsible humans for moral harm, victimized individuals (and those who care for them) may still seek out retributive justice and a place of refuge for the sense of outrage, anger, or other moral reactions.

### *Chapter 6 - Caring for Vincent: A chatbot for self-compassion*

As a moral emotion, compassion towards oneself can aid subjective well-being. Yet, increasing self-compassion via positive computing, i.e., technology for well-being, is underexamined. We hence looked into the relationship between the caregiver and care-receiver as human-computer interaction for self-compassion as a mixed-method study for two weeks. Specifically, while technology that guides people to care for themselves are well-established, we examined how people can care for a technological being as a way to care for themselves as a novel paradigm. We created a self-compassion chatbot (Vincent) and compared between caregiving and care-receiving conditions. Care-giving Vincent asked participants to partake in self-compassion exercises. Care-receiving Vincent shared its foibles, e.g., embarrassingly arriving late at an IP address, and sought out advice. While self-compassion increased for both conditions, only those with care-receiving Vincent significantly improved. In tandem, we shared qualitative data on how participants interacted with Vincent, e.g., giving compassionate advice to it. Our results demonstrated that when a person cares for a chatbot, the person's self-compassion can be enhanced. We further reflected on design implications for strengthening mental health with chatbots.

As above summaries of empirical chapters show, the presented re-

search is at the intersection of ethics, HCI, and other relevant disciplines. In what follows, I suggest areas for future research, offer critical remarks and broad insights. I end with elaborations on interactional morality.

One assumption that follows from the dissertation is that how people react to machines in dyadic moral interactions, i.e., descriptive acts, should inform normative positions. Or put differently, normative positions on how to design and use technology cannot be removed from what we descriptively do and can do. Though people may know they *should* treat machines as things, perhaps they *cannot help* but to treat machines as they would treat other people during morally relevant interactions. We then may benefit from examinations on when and how people's normative *ought*, e.g., "one should not treat a machine like a human", does not match their descriptive actions, e.g., treating it fairly and compassionately as one would treat a person.

When involving machines, the descriptive vs. normative divide may be differently drawn than in morally relevant human-human interactions. For instance, humans may blame and punish other people for wrongdoing, but perhaps only assign punishment (in whatever form), not blame, to artificial agents. Thus in prior chapters, I focused on three aspects: **(1) how people are affected by machines in morally relevant interactions, (2) if and when they cannot help but to treat machines as moral beings and (3) if and when they extend humanity to machines whilst also distinguishing themselves from machines.** In interactive contexts, people's behavior and their reflections on their behavior translate to the limits of what we can ask of people. HCI studies are recommended on limits of what people do vs. what they think they ought to do. This can help operationalize the descriptive vs. normative divide for our techno-moral futures, i.e., helping people do what they think they should do rather than what they normally would do.

There is evidence that suggests that we can extend our common humanity to artificial beings (Chapter 6), which can affect our moral emotions like compassion (Lee, Ackermans, et al., 2019). Participants showed signs of attachment to our chatbot Vincent. They made final queries like "can I keep him?" and did encourage Vincent while knowing that it was a fictional bot, e.g., "be proud of the bot that you are!" Simultaneously, people's self-compassion was influenced by a fictional bot. Whether the extension of our common humanity to technology has moral consequences, both short and long term, should be furthered researched.

I acknowledge that there is some danger to extending our humanity to artificial entities. Then are we better off not making bots that people can humanize (van Wynsberghe & Robbins, 2019)? Or is our moral circle already widening to include machines (Danaher, 2019) and should we welcome them with open arms? Yet what is truly new about this? The fact that ancient people prayed to the sun and that people marry an anime character in the present day (Jeffrey, 2018) both show that we have a longstanding tendency to expand our moral circle based on our needs, wants, and imagination. In what ways is it different *now* when it comes to artificial beings?

Our participants' behavior across chapters showed that indeed, our moral circle may be broadening, though not in consistent or predictable ways. People do not necessarily believe that machines can be blamed, but may still want to punish them in online contexts (Chapter 4). But, non-human agents are treated as if they truly deserve of compassion (like we deserve self-compassion) when they mirror our suffering (Chapter 5). Thus, what it means to humanize a machine or perceive a machine to be morally capable is complex. A nuanced view is that certain moral kinship is extended, e.g., experiences of compassion, whereas certain reactive attitudes and responsibilities are more likely reserved for humans only, e.g., blame. How we treat artificial agents differs from how we treat humans. And how we are affected by artificial agents differs from how we are affected by human beings. Still, these agents can be dyadic, conversational partners with which our emotions can be shared, even if they are not "one of us".

The dissertation focused on dyadic moral engagements through empirical investigations. Based on the results of presented studies, there are a number of considerations for understanding human-machine interactions of today and tomorrow. In Chapter 3, we have distinguished agency and patiency from cognition and emotion. While agency and patiency are intertwined conceptually, their distinction is significant within a dyadic interaction. A moral agent and patient are roles a person or a machine can have; one cannot be both at the same time in a single interaction (K. Gray, Young, & Waytz, 2012) though repeated interactions with switched roles are imaginable. In a dyad, each entity can be cognitively and emotionally well endowed, but can only act as a moral agent or a patient *in relation to the other actor*. Thus, if a person has a chance to act as a moral agent towards a machine patient, e.g., while negotiating or caring for its artificial suffering, then the machine is likely to be judged to have moral standing based on its perceived patiency, not agency.

Yet the roles of the agent and patient can change. Or, the agent and patient roles may evolve based on what the interaction paradigm allows. For instance, if the interaction paradigm invites a caring or compassionate stance, the caregiver and the cared-for can take turns caring for one another (Noddings, 2008). Thus, who becomes a moral agent or patient depends on the moral elements of an interaction. In practical terms, what this means is that it is never *just* about an AI system or a chatbot being moral or immoral, morally capable or incapable. Instead, a human interactant can render (or allow) the technological being to be a moral agent or patient within an interaction, given what a person can do *against* it, e.g., negotiate a better offer, or what they can do *for* it, e.g., offer an uplifting advice to make a chatbot feel better, within situational bounds.

Interactants typecast one another (as an agent or patient) given the situation they are in. Hence, even if a machine's moral capacity seems "ready-made" to humans as algorithmic calculations that support an ethical theory, the built-in capacity is not what makes it moral. What makes a machine seem moral to human interactants is how its "moral processing" has the chance to be expressed towards people. Then the perception that the machine is moral can hinge on people's perception and standpoint, such as whether or not it deserves to be protected from harm, or whether or not it subscribes to one's preferred ethical theory.

People were able to express their preferred moral code of conduct for studies in Chapters 4 and 5 with a robot on the trolley dilemma and footbridge dilemma. Even if people disagreed with the robot, participants' moral position (though limited to deontology vs. utilitarianism) did not sway their moral decision in Chapter 4. People saw the robot as an amoral agent because it had no emotions, i.e., it can make moral decisions due to its calculated, rational decision-making, but this behavior is not considered to be moral. Additional environmental elements, like the visual transparency cues, were then likely to be attributed as enhancing a machine's cognitive and agentic traits according to our quantitative analysis. Based on qualitative data, we saw that machines' lack of emotions was one prominent factor in how participants distanced themselves from a robot; since a robot cannot feel the emotional intensity of harming someone, it cannot truly know the weight of causing death like humans can.

In Chapter 5, we noted that people had a tendency to attribute greater agency than patiency to a robot (in line with literature (H. M. Gray et al., 2007)). A robot that showed emotional displays was attributed

with a higher level of agency than a robot that did not show emotional displays; in fact, even for a robot that behaved emotionally, its perceived agency was consistently higher than perceived patiency across three studies (Chapter 5, Figure 5.2). Chapter 5 also demonstrated how the lack of perceived emotions in a robot may lead people to consider punishing it, but not blaming it, much like retributive justice for psychopaths (Godman & Jefferson, 2017). This is in line with our qualitative data in Chapter 4 that showed emotions to be critical in moral decision-making.

Rather than designing and researching on how machines can truly appear to be emotional, a path we took was on amplifying human emotions by giving people opportunities to be compassionate towards a machine as a way to be self-compassionate (Chapter 6). Neither our real emotions need to be untainted by artificial emotions nor is this conceived to be possible— we share emotions with other people without needing or requiring clear ownership of felt emotions, so people can build on or be affected by artificial feelings (Chapter 2). As long as a an artificial agent's identity as a mere machine was transparently communicated to involved parties, sharing of emotions was conceivable and acceptable, especially if real and artificial emotions together can contribute to better understanding and shaping of oneself (Chapter 2). Yet an interesting contrast is found between Chapter 4 and 5: a robot is perceived to be unable to make moral decisions with emotions (Chapter 4), but people may be more likely to punish a robot that does not display emotions compared to a robot that displays artificial emotions (Chapter 5).

While expecting machines to be non-emotional, people's emotions can still be influenced by a robot's display of artificial emotions. Hence, we dived into one moral emotion in Chapter 6, i.e., compassion as a positive moral emotion or motivation (Haidt, 2003). Compassion was tied to perceived patiency of an artificial agent in contributing to people's change in reported compassion states, as exploratory findings (Chapters 3 and 5). We saw that self-compassion can increase even with a single interaction of ten minutes if people can discuss non-trivial life happenings, like the most recent memory of failure (Chapter 6). In this, the identity of a machine partner seemed to matter less in one-shot interactions as long as people can be vulnerable, e.g., recounting failure (Chapter 6). How a machine talks or acts, i.e., its identity, does factor in more for long-term engagements (two weeks or longer). In a common sense way, Vincent studies demonstrated that the longer one gets to know a machine, the way it behaves may influence how people feel more than when people have a single encounter with it.

There are limitations and potential areas for future work. Studies were exploratory that should be followed up with more substantial efforts, both qualitatively and quantitatively. Further, neither all morally relevant constructs were looked into nor was one specific construct repeatedly investigated. For example, mind perception between moral dyads figure into compassion theoretically, but how the bridge could be made empirically remains to be better understood. Efforts should be made on if and how machines that divert from people's moral positions could play a role in everyday life. Of moral emotions, only compassion was included and whether other moral emotions like guilt or shame could be emotions that impact people through machines could be additionally explored. The human-machine dyad has been the focus thus far, but how moral dimensions of group-level dynamics could develop is an area to delve into. The last issue is that there are myopic considerations on what ethical traditions are given attention, e.g., deontology over compassion, in shaping the current discourse on morality. The current dissertation does not go beyond or more critically look into different ethical traditions, especially from more varied cultures.

### 7.3 Setting the scene

Here, I zoom out from prior chapters to establish further theoretical connections on interactional morality. To begin, I elaborate on the concept of human-likeness of machines. It is well-established that we treat machines socially as we would treat other people as the Computers Are Social Actors (CASA) paradigm suggests, e.g., people greeting computers when this behavior is not required for machines (Fogg & Nass, 1997; Moon & Nass, 1996; Nass et al., 1994). So we greet people, but also computers *as if* they are like people. What about computers that look like humans? Human-computer interaction (HCI) research has long been concerned with *human-like* machines, e.g., the Uncanny Valley (Mori, 1970). Yet, "human-like" machines often are typecast as literal humanoid figures that look and talk like us, e.g., as a virtual human therapist, rather than a more nuanced research that focus *only* on human-like *behavior*. When agents are made to look like us, we see that human biases based on appearance, e.g., gender or race of virtual people, creep up in how these artificial entities are judged (Bailenson et al., 2003; Dotsch & Wigboldus, 2008; Ruijten et al., 2015; Siegel et al., 2009). This indirectly touches on morality, namely how our human world biases become present in the virtual world when machines *look* like us.

However, a greater consideration can be on what exactly we want from machines compared to what we can expect from them. When people feel that a virtual human (Lucas et al., 2014) and chatbot (Brandtzaeg & Følstad, 2017) are less judgmental compared to other humans, the embodied or non-embodied nature of artificial agents seem to be of secondary importance. The categorization of a machine as a virtual agent or robot may not be so important to people if what they are looking for is a non-judgmental machine. The CASA paradigm (Fogg & Nass, 1997; Moon & Nass, 1996; Nass et al., 1994) indeed shows that we can socialize with machines, but the deeper thinking is on *why we want to socialize with machines at all*, superficially or not. What underlies the perception that machines are non-judgmental points to our fear of being judged by other humans. Our fear of being judged by humans, then, is on how we imagine other people to perceive us. Artificial agents as “mere machines” cannot perceive and judge us like other humans can (Lucas et al., 2014).

We can design machines’ perceptions of us more readily than other people’s perception of us. As a simple example, a robot that interactively gazes at a talking person has been designed to detect and follow the movement of a human face and body. Compared to humans, machines seem to offer a truly blank slate on who we are seen to be, whereas people we meet could judge us based on our gender, race, among other traits. With other people, morally relevant interactions often seem to hinge on these perceived trait-based biases. There are many examples, but most recent ones in my time of writing would be selective police brutality towards Black Americans.<sup>1</sup> Indeed, human-human moral interactions come with many human biases based on our external features. And if these biases and perception of one another are a part of our experiences of human morality, what are we seeking in “human-like” machines when transitioning to moral HCI? What should our focus point be?

We hope that we are not treated differently based on our gender, skin color, accents, physical abilities, among other features; we also hope that we do not treat others differently based on their gender, skin color, accents, physical abilities, among other features. We hope that our *behavior* dictates how we see ourselves to be morally good or immoral, rather than our perception of our own and others’ gender, skin color, accents, physical abilities, among other features. When focusing solely on behavior, whether we treat machines in a moral manner (and what that means) from an interactive, first-person perspective is only beginning to be explored.

<sup>1</sup> Star Tribune:  
<https://www.startribune.com/minneapolis-police-marchers-clash-over-death-of-george-floyd-in-custody/570763352/>. A specific case is a white CNN reporter not being arrested in the same premise during riots in Minneapolis, Minnesota, while his Black-Latino CNN colleague was arrested by the police. CNN: <https://edition.cnn.com/2020/05/29/us/minneapolis-cnn-crew-arrested>



Studies that involve people as first-person interactants who experience a moral situation with machine can aid what ethics can become for our times ahead. What I think about a robot or how I feel about a chatbot is likely to differ based on whether I directly interact with them or indirectly judge them from the sideline— as humans, being a part of a social drama as it unfolds for oneself is vastly different from watching others undergo such drama. Thus, I point out the importance of *vantage points* that should incorporate two overlooked aspects: (1) a partial, first-person perspective rather than an impartial, third-person perspective and (2) how the first-person perspective is *activated* and *changed* in morally relevant interactions. I sketch out the gap in research with a few examples and issues to be more deeply addressed in the future.

One major issue is on the normative portrayal of anthropomorphism and mind. Scales of “human-likeness” or perceived mind used in HRI (Carpinella et al., 2017; H. M. Gray et al., 2007; K. Gray & Schein, 2012; Waytz, Morewedge, et al., 2010) assume psychological normativity and thereby exclude people with abnormal psychology, such as sociopaths or psychopaths who may lack greater emotional sensitivity, though they are humans. As such, when a robot is assessed to be “human-like”, this judgment implies a comparison to a psychologically and developmentally “normal” human adult when it could also be as human-like as a psychopath. There are other related issues on biases such as gender typecasting of concepts, i.e., competence being more related to males and warmth to females (Carpinella et al., 2017). As a different approach, we can conceptualize how well robots can be human-like by their ability to surprise us, possibly disambiguating assumptions underlying artificial intelligence to include unpredictability, creativity, emotional complexity, as well as logical cognitive processing (refer to the Lovelace test (Bringsjord, Bello, & Ferrucci, 2003)), which are all different aspects of being a person.

To make further progress, broadening anthropomorphism at the core can start by confronting what norms we uphold or disregard when we call technology “human-like”. Not only can moral human-machine interactions contribute to how and why ethics is integral to AI for our society, it can help us reflect on our moral opinions and biases as individuals. Furthermore, observing how we interactively form our opinions on whether robots have independent, artificial minds can expand our own minds (A. Clark & Chalmers, 1998). As researchers, we can then reconsider the normative assumptions of a human-like mind that often translate to the design of artificial minds without in-depth considerations of our biases on what counts as being human or having

a mind. When we inherit norms on anthropomorphism through the scales used in HRI research without explicating underlying assumptions, we miss the chance for a deeper exploration on how divergent humans can be, and in turn, how diversely robots can be human-like.

There are critical issues with the way common scales we have adopted were created. We covered that mind perception of machines is key in how humans and machines interact with one another (Krämer, 2008). Yet, the Mind Perception Theory (MPT) is built on participants comparing different beings, like robots and human adults, to each other from an impersonal, third-person perspective (H. M. Gray et al., 2007; K. Gray, Young, & Waytz, 2012). MPT was not designed with the first-person perspective as a priority. In applying MPT in empirical contexts, people are asked to form opinions about actors in *hypothetical* scenarios as impartial *third-party* judges (K. Gray & Wegner, 2012). In moral psychology, third-person perspectives are also common. For example, a famous case is a fictional story on a character named Heinz and whether he should steal an unaffordable drug to help treat his partner's cancer treatment; such stories were deployed to judge people's level of moral development (Kohlberg, 1969, 1973).

Thus, the assumption is that people's moral progress can be based on how they reply to hypothetical situations without themselves being directly integrated. In human-computer interaction (HCI), third-person perspectives were also deployed to study morality. We see experiments that used scenario-based moral dilemmas about robots. Though the topic of the experiment was explicitly about morality (the trolley dilemma), participants were asked to be third-party judges of robots' actions, i.e., participants did not directly interact with a machine (Komatsumi, 2016; Malle et al., 2015).

First-person interactions covered in HCI studies on negotiations, games, or economic exchanges contain morally relevant features *implicitly* by looking into fairness; prior studies showed that a machine's display of emotional expressions, as well as framing of whether the machine is controlled by a human or is autonomous, affects people's tendency to cooperate with it or treat it fairly (Baarslag et al., 2017; de Melo, Gratch, & Carnevale, 2014; de Melo et al., 2018; de Weerd et al., 2017; Veltman et al., 2019). Yet, caveats arise. As mentioned, moral elements were only implicit in these studies, not an explicit topic of investigation. More broadly, fairness is only one aspect of morality (Graham et al., 2011). Even then, HCI studies do not yet cover the human *self* as a nuanced moral being.

One concept that illustrates the nuanced self is *moral self-deception*: how people perceive that they acted fairly while in practice, they acted unfairly. In a series of studies on fair allocation of tasks, people opted to appear fair without the cost of being fair, i.e., flipping a coin as an act of being fair even though they disregarded the coin toss results and allocated better tasks to themselves than others; the self-perception of being a fair person was based on symbolically flipping a coin even if it did not affect the end decision (Batson, Kobrynowicz, Dinnerstein, Kampf, & Wilson, 1997; Batson, Thompson, & Chen, 2002; Batson et al., 1999).

Another example is *moral disengagement*: how people disregard or amoralize (make morally irrelevant) ethical factors of a situation (Bandura, 1999, 2016; Bandura, Barbaranelli, Caprara, & Pastorelli, 1996). For instance, executioners working at prisons were found to make moral aspects of their job mundane, like dehumanizing prisoners in order to believe that “we have a job to do and that job isn’t to be a coldhearted individual. It is simply to carry out the order of the state”, according to an anonymous executioner (Osofsky, Bandura, & Zimbardo, 2005). Thus from flipping a coin to one’s benefit to *normalizing* moral cruelty, many of us do morally disengage to maintain a positive self-image or what should be considered morally wrong is made mundane— what Arendt called “banality of evil” (Arendt, 2006 [1963]). There are many sides to *being human* that HCI research on morality should begin to address.

Lastly, a greater diversity on ethical theories or frameworks that serve as backdrops to understand moral HCI is necessary. The popular trolley dilemma (Foot, 1967), that I am guilty of deploying here, supposedly contrasts deontology and utilitarianism, as if they are the only ethical theories that people subscribe to and as if within each tradition, there is no diversity of philosophical views. And as researchers, especially in the quantitative tradition, we lack the means to treat participants as moral beings on their own right who may express moral sentiments and thoughts that are diverse, and perhaps not aligned with either deontology or utilitarianism. Then the machine’s design fares no better. The bottom-up or top-down training of a machine, or even their hybrid approach (Wallach & Allen, 2008) is likely to be too vague for situational decision-making; humans widely vary on what ethical route to take (if there is even time to ponder on this) per scenario (Darley & Batson, 1973; Doris, 1998). Thus a machine’s moral processing of our ethical theories is likely to run into the same situational nuances we face in real life.

In sum, though many works implicitly hint at morality, less HCI research is explicitly about moral interactions. And if they are explicitly about moral interactions, many are on indirect third-party judgments of presented scenarios, not one-on-one interactions with technology. Further, the richness of how nuanced the moral *self* can be, e.g., practicing moral self-deception or moral disengagement, should begin to be better explored in empirical settings. The path forward can benefit from a greater diversity of ethical traditions for grounding empirical and philosophical works, as well as our willingness to consider that people's notion of morality may not fit well-established paradigms. To do so, interactional morality attempts to understand people's perspectives as moral beings who undergo morally pertinent experiences with second-person artificial beings.

## 7.4 Interactional morality

Often, our world does not tidily present us with morally good and immoral choices *in situ*. Moral conflicts on what is the right action to take can be daunting (Tessman, 2017). A recent example is on healthcare professionals who could not save all lives during emergency situations and made difficult choices on who to save during the coronavirus pandemic of 2020.<sup>2</sup> Hopefully less dramatically, most of us had to make moral choices as events unfolded in front of us, such as the choice to speak up as children in school playgrounds or as adults during work meetings. In all morally relevant situations, there is always an interplay between actors and the situation they find themselves in. Having machines join us in our everyday settings complicates morality for all of us with potentially new breeds of moral conflicts and decisions.

In ethics of technology and moral HCI, *interactional morality* stands for an interplay between three elements: a *person* and a *machine*, within a particular *situation*. The moral relevance can stem from the situation, e.g., emergency healthcare response, the person, e.g., experiencing moral emotions like compassion or guilt, or the machine, e.g., algorithmic emotion display. Importantly, interactional morality emphasizes that putting our efforts in each factor individually would not help us going forward. Designing machines to be ethical (if possible) (Moor, 2006; Wallach & Allen, 2008) would not alone provide a solution and potentially reduce post-hoc accountability by opening up responsibility and retributive gaps (Danaher, 2016; Matthias, 2004; Nyholm, 2018; Sparrow, 2007). Relying on our moral character (Aristotle, 2011 [±340 BC]) alone does not translate to how human morality may

<sup>2</sup> The Washington Post - Spiking U.S. coronavirus cases could force rationing decisions similar to those made in Italy, China <https://www.washingtonpost.com/health/2020/03/15/coronavirus-rationing-us/>

morph with technology (Vallor, 2016). Focusing only on the situation itself (Darley & Batson, 1973; Doris, 1998) may make many of us feel powerless to enact change.

Interactional morality is a conceptual contribution in HCI. There are many predecessors, though not all can be covered here. As explained below, I build on Kurt Lewin's contribution (1931) as extended by Daniel Batson on morality (2017) as a starting point. What is further added is the emphasis on the experience morality as foremost *interactive* (Darwall, 2004, 2006). Based on interactions, we blur self-other distinction between humans, e.g., in practicing compassion (Breines & Chen, 2013; Kongtrul, 1987; Neff, 2003a; Tse, 2008). The blurring self-other distinction, then, also may occur between humans and machines in the case of compassion and self-compassion (Falconer et al., 2016; Lee, Ackermans, et al., 2019).

Lewin's now taken for granted dictum is that psychology is not just about a person, but the person-situation interaction in understanding a person's behavior. Ergo, a person's *behavior* (B) is a function (f) of the person (P) and the environment or situation (E), represented as  $B=f(PE)$  (Lewin, 1951). Lewin's earlier work points to a deeper insight: there is conflict between Aristotelian and Galilean conceptualization of science that pervades much of mainstream psychology (1931), even to this day. Aristotelian science classified observations to establish *phenotype* features. For instance, objects that are light tend to float so they have a tendency to go upward, whereas objects that are heavy sink, so they have a downward tendency, in Aristotle's scientific world view (Lewin, 1931). Thus, floating or sinking was seen to represent *intrinsic* properties of objects themselves. However in the Galilean world view, there are *extrinsic* forces at play, such as the relationship between gravity and objects themselves. We cannot literally see gravity, but it underlies concepts like acceleration of objects. So Galilean science focused on the *genotype* that provide a more comprehensive explanation rather than phenotype classifications.

Our planets orbiting the solar system and leaves falling seem like disparate behaviors until gravity enters the picture (Lewin, 1931). Hence, genotype allows for exceptions to be explained within a cohesive system whereas phenotypic thinking would not categorize exceptional observations alongside frequently occurring behaviors. Lewin's remark is that still, we tend to think of a person and their attributes in a phenotypic manner in modern psychology. For instance, frequent observations of when a person was behaving fairly would seem to represent that that person is fair, in which unfair acts would be exceptional.

Fairness, then, would be a person's trait and a virtuous person would possess many noble traits, like courage, in an Aristotelian sense (Aristotle, 2011 [ $\pm 340$  BC]; Lewin, 1931), which is a phenotype approach. What could instead be a genotypic starting point?

Lewin proposed that *motives*, *values*, and *goals* can better explain human *behavior* than traits or character (Lewin, 1931)— so the focus is shifted from being fair and how a person developed to be fair, to acting fair based on the *situational* interplay of a person's motives, values, and goals, for instance. Values here can be everyday pleasures like enjoying a cup of tea to more abstract concepts like fairness and justice (Friedman, 1996; Friedman et al., 2008); goals are determined when opportunities to gain and threats to lose what one values arise; motives represent the *how* of pursuing a goal, or goal-directed movement (Lewin, 1931). We are not always consciously aware of our motives, values, and goals (Batson et al., 1997; Lewin, 1931).

While character cannot fully explain behavior, particularly moral behavior, one's character is undeniably important, especially our need to see ourselves in a positive moral light (Batson, 2017). Aristotelian virtues that are the basis of moral character (Aristotle, 2011 [ $\pm 340$  BC]) can be accommodated in Lewin's framework when considering how one's values come to be held (Batson, 2017). Can a moral value be intrinsically held as good in and of itself? Yes and that is the aim for Aristotle as well. But how is any value *enacted* as behavior? The gap that Batson addresses is that it is rare to act *solely* out of internalized value that is integrated into one's behavior. Rather, people most often act to be judged as morally good in their own and others' eyes (Batson, 2017). Then, avoiding immoral actions is due to fear of punishment or loss of face, i.e., introjecting external norms or rules to regulate one's behavior (Batson et al., 1997; Deci, Eghrari, Patrick, & Leone, 1994). Hence, the significance is that (1) the lack of immoral behavior does not signal moral behavior, but also that (2) morally good behavior based on introjection (internalized external criteria, e.g., rules) may neither be authentically good nor truly moral. Batson's conclusion is that *appearing* moral to oneself and others is far more common than abiding by values that are integrated to one's behavior (2017).

When, however, is it possible to detect that values have been sufficiently integrated into one's behavior? Again, in the current norms of empirical research, we are likely to fall prey to counting frequencies of behavior to remark about character. Whether or not we abide by values in our daily acts may be more of an aspirational question than an empirical question. What can be gained from prior research are

tactics that can support one's aspiration to do better, while somehow avoiding self-deception and moral disengagement in order to see oneself to be good, regardless of one's actions. It almost seems impossible for most of us; even Aristotle's conception was that ethics, as he saw it, was not meant for everyone (Aristotle, 2011 [ $\pm 340$  BC]). What can be done?

A path is on activating awareness of one's goals, values, and motives. If one values acting in a fair way rather than being perceived by others to be fair, a motivating element is a reminder for fair behavior. In a series of studies on moral self-deception, one manipulation strongly lowered people's likelihood of pursuing self-beneficial choices. When people were flipping coins while facing a reflective mirror, the chance that they would assign themselves with better tasks was 50% regardless of the coin toss (Batson et al., 1999)—significantly lower and more realistic than in studies without the reflective mirror, in which above 80% of participants granted better tasks to themselves regardless of the coin toss that was meant to decide who gets which tasks (Batson et al., 1997, 2002, 1999). Technology could act as a literal mirror; we can even take on different perspectives that we would be unable to in reality, e.g., through VR (Falconer et al., 2016). Via interactional morality, technology could thus remind us of our goals, values, and motives, not just for fairness, but also compassion.

## 7.5 Merging of the self and the other: Compassion

Interactional morality involves *oneself* interacting with a *machine* within a particular *situation*. To better assess the self within a situation, I above elaborated on focusing on people's goals, values, and motives rather than steadfast character. Here, we consider how people see themselves in others (both human and machine others) via compassion. While seeing oneself in other people as shown by behaviors like looking up to someone or commiserating with another may be a common experience, seeing oneself in a machine is unintuitive for most of us. We explored compassion empirically to assess boundaries of who we can be compassionate towards, possibly even machines, as a means to be compassionate to ourselves (Chapter 6).

What is intrinsic to compassion is the lowering of self-other distinction. When people are asked to practice perspective-taking or to see themselves in another person, people can infuse others' freely chosen acts as attributes of one's own behavior, extending the limits of selfhood

(Goldstein & Cialdini, 2007). Similar merger happens in compassion, which is why it is distinct from emotions such as sympathy, empathy, grief, or pity (Cartwright, 1988; Nichols, 2004; M. Nussbaum, 1996). Compassion offers an ego-less perspective through shared suffering and provides a moral orientation, i.e., the greater humanity is merely “‘myself once more’” (Schopenhauer, 1995 [1840], p. 277).

Other’s suffering can cause empathic distress— a helpless, negative feeling of vicariously experiencing others’ suffering, rather than activating altruistic motivation underlying compassion (Calvo & Peters, 2014; Nichols, 2004). Hence empathy can have a “dark” side through distress at another’s distress (Haidt, 2003), which targets processing of suffering different way than compassion (Bloom, 2017). Compassion is a unique moral emotion that may be more helpful than empathy because suffering underlying compassion is a complex experience that cannot be easily labeled as positive or negative. Some can grow because of suffering while others may wish to distance themselves from it. And to suffer can be to cherish what it means to be human from one angle (Nietzsche, 1989 [1885]). Within oneself, one’s relationship to suffering can fluctuate over time; we can learn to have a more constructive relationship by witnessing our own and others’ suffering through compassion.

Unlike empathic relatedness to others that comes more automatically that can cause empathic distress, compassion requires a training process. In one study, when people were exposed to others’ suffering, compassion training showed to enhance their positive affect, counter negative affect, and decrease empathic distress, compared to empathy training that only activated negative affect and empathic distress when looking into neural mechanisms, i.e., the pain circuitry (Klimecki, Leiberg, Ricard, & Singer, 2013). When operationalizing compassion in mental health endeavors, the ability to conceptualize and emotionally relate to suffering matters the most, and this relatability can nurture compassion (Tse, 2008).

In applied ethics involving technology, compassion does not commonly fit into mainstream ways of merging HCI and ethics. But tides may be changing. Positive computing does give room for compassion in how to conceptualize well-being with technological means for the present age and beyond (Calvo & Peters, 2014). In positive computing, we see that compassion starts with the self (Calvo & Peters, 2014), but a nuanced view on the self is lacking. We may act in ways that do not align with our moral principles, yet we easily avoid dissonance via making exceptional cases so that our self-narrative remains largely un-



changed. It is difficult to be aware of inner inconsistencies on what we morally believe versus what we end up doing, with the remembering self and experiencing self sometimes being two very separate individuals (Kahneman & Riis, 2005). The human-like AI we are envisioning when we entertain how technology can be as moral as us may first require more reflection on the varied ways we carry out our motives, values, and goals. Perhaps the vulnerable, moral individual can face AI, not as an algorithm in the background, but as a mirror that shows one's own moral compass to oneself. Who we see ourselves to be and how we allow ourselves to change is an ethical region of interest that HCI can better address.

There is a possibility to start on a different footing when we notice how dynamic an interaction between any human and machine can be (Šabanović, 2010), but also how much our self-perception can change along with technology we interact with. Accepting the dynamic nature of any moral interaction helps us to see technology as extensions of who we are; they are already embedded in our daily routines and envisioned selves (A. Clark & Chalmers, 1998; Turkle, 2016; Vallor, 2016). It is not a huge jump to imagine that they extend our ethical compass when we face technology as our vulnerable selves. Yet asking technology to reveal if we are only self-deceptively "fair" might not be what we want to see.

We learn to how to become attentive to morally relevant situations over time and figure out ways to respond to them in a variety of ways. For instance, we may perhaps be motivated by compassion, deontological ethics, or even through moral self-deception. Most of us want to maintain a positive self-image. To assist, an interactive machine can be individualized with a narrative of its own (Sengers, 2000) to show us with a gentler, less judgmental process to uncover e.g., self-deceptive fairness. Rather than telling someone that they cheated and that goes against their principle of fairness, we could show someone that another being they can relate to, machine or human, cheated, but realized that action was not fair. Rather than telling someone that they should be less judgmental towards themselves, showing another being, machine or human, becoming less self-judgmental could mirror the process of oneself becoming less judgmental. These methods can capitalize on the self-other merger or ways in which we find greater humanity in others. Perhaps we could learn to be aware of our morally good and immoral sides (as well as everything in between) of who we are.

Morality starts with our particular selves with our partial ways of be-

ing, existing, and seeing the world. Just as we want to be treated as individuals and not be tokenized by others, there is danger in tokenizing others as technological, organic, or anywhere in between, beings: "Kindness occurs best when one is not tokenizing the recipient of one's kindness, since only then can one tailor one's kind acts to the particular needs of an individual" (Tse, 2008), including the particular values and experiences of (and for) our individual selves. In many ways we have tokenized machines as clunky, metal robots or as algorithmic, "black box" mysteries. People believe that machines can be logical, but not emotional (H. M. Gray et al., 2007), though machines can neither be logical nor emotional in the same way as humans can be. Importantly, our beliefs about an artificial agent (like another person) can be overturned or recalibrated *within and through a situation* based on our own ethical vantage points, which affects our view of our moral selves.

Interactional morality suggests that one's morally relevant characteristics (like agency) require, at minimum, a recipient being or thing; the capacity to be agentic is best demonstrated as an interaction towards another being or thing in a particular situation. Likewise, compassionate feelings are demonstrated towards a being or thing within a particular situation. Further, interactional morality includes how I can interact with my perception of who I am through technology, e.g., as moral feedback. An interaction can be between oneself and one's reflected self in dyadic technology as a particular situation.

In sum, stories of tokenized machines as the clunky other can be revised if we see ourselves reflected in technologies we use and interact with. For example, technology is not emotional like we are, but our emotional experiences can be enriched by our interactions with various artificial agents that are perceived to have human-like feelings. Technology that extends our self-perception can also help us retell who we are to ourselves, e.g., as beings deserving of self-compassion rather than self-judgement. The question is on what stories we choose to tell and hear in the moral mirror. Artificial agents' partial ways of being, existing, and seeing the world are our own to explore and nurture.



## References

- Ackerman, M. S. (2000). The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2-3), 179–203.
- Admoni, H., Datsikas, C., & Scassellati, B. (2014). Speech and gaze conflicts in collaborative human-robot interactions. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36).
- Admoni, H., & Scassellati, B. (2017). Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1), 25–63.
- Adorno, T. W. (2005 [1951]). *Minima Moralia: Reflections from Damaged Life*. London, UK: Verso.
- Alicke, M. D. (2012). Self-injuries, harmless wrongdoing, and morality. *Psychological Inquiry*, 23(2), 125–128.
- Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3), 149–155.
- Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019)* (p. 1078–1088). Richland, SC, USA: International Foundation for Autonomous Agents and Multiagent Systems.
- Anxiety and Depression Association of America (ADAA). (2016). *ADAA Reviewed Mental Health Apps*. <https://adaa.org/mental-health-apps>. ADAA. (Accessed: 2018-06-25)
- Arditte, K. A., Çek, D., Shaw, A. M., & Timpano, K. R. (2016). The importance of assessing clinical phenomena in Mechanical Turk research. *Psychological Assessment*, 28(6), 684.
- Arendt, H. (2006 [1963]). *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York, NY, USA: Penguin.
- Aristotle. (2011 [±340 BC]). *Aristotle's Nicomachean Ethics*. Chicago, IL,

- USA: University of Chicago Press. (Translated by Bartlett, Robert C. and Collins, Susan D.)
- Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction (HRI 2008)* (p. 121–128). New York, NY, USA: Association for Computing Machinery.
- Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4), 596.
- Averill, J. R. (1980). A constructivist view of emotion. In R. Plutchik & H. Kellerman (Eds.), *Theories of Emotion* (pp. 305–339). New York, NY, USA: Academic Press.
- Baarslag, T., Kaisers, M., Gerding, E., Jonker, C. M., & Gratch, J. (2017). When will negotiation agents be able to represent us? The challenges and opportunities for autonomous negotiators. In S. Carless (Ed.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4684–4690).
- Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. *Personality and Social Psychology Bulletin*, 29(7), 819–833.
- Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *2008 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 701–706).
- Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193–209.
- Bandura, A. (2016). *Moral Disengagement: How People Do Harm and Live with Themselves*. New York, NY, USA: Macmillan.
- Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, 71(2), 364–374.
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1), 1–68.
- Barry, B., Fulmer, I. S., Van Kleef, G. A., et al. (2004). I laughed, I cried, I settled: The role of emotion in negotiation. In M. J. Gelfand & J. M. Brett (Eds.), *The Handbook of Negotiation and Culture* (pp. 71–94). Stanford, CA, USA: Stanford Business Books.
- Bartneck, C., & Keijsers, M. (2020). The morality of abusing a robot. *Paladyn, Journal of Behavioral Robotics*, 11(1), 271–283.
- Bates, J. (1994, July). The role of emotion in believable agents. *Commun. ACM*, 37(7), 122–125.

- Batson, C. D. (2008). Moral masquerades: Experimental exploration of the nature of moral motivation. *Phenomenology and the Cognitive Sciences*, 7(1), 51–66.
- Batson, C. D. (2017). Getting cynical about character: A social-psychological perspective. In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral Psychology: Virtue and Character* (Vol. 5, p. 11). Cambridge, MA, USA: MIT Press.
- Batson, C. D., Kobryniewicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: Unmasking moral hypocrisy. *Journal of Personality and Social Psychology*, 72(6), 1335–1348.
- Batson, C. D., Thompson, E. R., & Chen, H. (2002). Moral hypocrisy: Addressing some alternatives. *Journal of Personality and Social Psychology*, 83(2), 330–339.
- Batson, C. D., Thompson, E. R., Seufferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: Appearing moral to oneself without being so. *Journal of Personality and Social Psychology*, 77(3), 525–537.
- Baudrillard, J. (1994). *Simulacra and simulation*. Ann Arbor, MI, USA: University of Michigan Press. (Translated by Sheila Faria Glaser)
- Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8(9), 536–554.
- BBC. (2019, 8). *Why I 'married' a cartoon character*. <https://www.bbc.com/news/stories-49343280>, note = Accessed on 01/24/2020, publisher = BBC News.
- Beck, J. (2013, September). *Married to a doll: Why one man advocates synthetic love - The Atlantic*. <https://www.theatlantic.com/health/archive/2013/09/married-to-a-doll-why-one-man-advocates-synthetic-love/279361/>. (Accessed on 01/25/2020)
- Bentham, J. (1996 [±1789-1843]). *The Collected Works of Jeremy Bentham: An Introduction to the Principles of Morals and Legislation* (J. Burns & H. Hart, Eds.). Oxford, UK: Oxford University Press.
- Bergman, R. (2004). Identity as motivation: Toward a theory of the moral self. In D. K. Lapsley & D. Narvaez (Eds.), *Moral Development, Self, and Identity* (Vol. 2, pp. 21–46). Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.
- Black, J. E., & Reynolds, W. M. (2016). Development, reliability, and validity of the moral identity questionnaire. *Personality and Individual Differences*, 97, 120–129.
- Blackburn, S. (2002). *Being Good: A Short Introduction to Ethics*. Oxford, UK: Oxford University Press.
- Blair, R. J. R. (1997). Moral reasoning and the child with psychopathic

- tendencies. *Personality and Individual Differences*, 22(5), 731–739.
- Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2), 103–124.
- Blasi, A. (1993). The development of identity: Some implications for moral functioning. In G. G. Noam, T. E. Wren, G. Nunner-Winkler, & W. Edelstein (Eds.), *The Moral Self* (pp. 99–122). Cambridge, MA, USA: MIT Press.
- Blasi, A. (1999). Emotions and moral motivation. *Journal for the Theory of Social Behaviour*, 29(1), 1–19.
- Blasi, A. (2004). Moral functioning: Moral understanding and personality. In D. K. Lapsley & D. Narvaez (Eds.), *Moral Development, Self, and Identity* (Vol. 2, pp. 335–347). Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.
- Bloom, P. (2017). *Against Empathy: The Case for Rational Compassion*. London, UK: Random House.
- Blythe, M. (2017). Research fiction: Storytelling, plot and design. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 5400–5411). New York, NY, USA: Association for Computing Machinery.
- Bohnet, I., & Frey, B. S. (1999). Social distance and other-regarding behavior in dictator games: Comment. *American Economic Review*, 89(1), 335–339.
- Borenstein, J., & Arkin, R. (2016). Robotic nudges: The ethics of engineering a more socially just human being. *Science and Engineering Ethics*, 22(1), 31–46.
- Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In I. Kompatsiaris et al. (Eds.), *Internet Science* (pp. 377–392). Cham: Springer International Publishing.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Braun, V., & Clarke, V. (2012). Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA Handbook of Research Methods in Psychology. Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological* (pp. 57–71). Washington, D.C., USA: American Psychological Association.
- Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-computer Studies*, 62(2), 161–178.
- Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications*

- and Reviews*), 34(2), 181–186.
- Breines, J. G., & Chen, S. (2013). Activating the inner caregiver: The role of support-giving schemas in increasing state self-compassion. *Journal of Experimental Social Psychology*, 49(1), 58–64.
- Bringsjord, S., Bello, P., & Ferrucci, D. (2003). Creativity, the Turing test, and the (better) Lovelace test. In J. H. Moor (Ed.), *The Turing Test: The Elusive Standard of Artificial Intelligence* (pp. 215–239). Dordrecht: Springer Netherlands.
- Broadbent, E., Tamagawa, R., Kerse, N., Knock, B., Patience, A., & MacDonald, B. (2009). Retirement home staff and residents' preferences for healthcare robots. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 645–650).
- Brooks, D., Shultz, A., Desai, M., Kovac, P., & Yanco, H. A. (2010). Towards state summarization for autonomous robots. In *2010 AAAI Fall Symposium Series*.
- Brownlee, K. (2013). A human right against social deprivation. *The Philosophical Quarterly*, 63(251), 199–222.
- Brownlee, K. (2014). Ethical dilemmas of sociability. *Utilitas*, 28(1), 1–19.
- Bühler, M. (Ed.). (2015). *No Internet, No Art: A Lunch Bytes Anthology*. Eindhoven, the Netherlands: Onomatopoe.
- Buzan, T., & Buzan, B. (2006). *The Mind Map Book*. London, UK: Pearson Education.
- Calvo, R. A., & Peters, D. (2014). *Positive Computing: Technology for Wellbeing and Human Potential*. Cambridge, MA: MIT Press.
- Cardon, A. (2006). Artificial consciousness, artificial emotions, and autonomous robots. *Cognitive Processing*, 7(4), 245–267.
- Carey, M. A., & Smith, M. W. (1994). Capturing the group effect in focus groups: A special concern in analysis. *Qualitative Health Research*, 4(1), 123–127.
- Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83(2), 284.
- Carlsson, A. B. (2017). Blameworthiness as deserved guilt. *The Journal of Ethics*, 21(1), 89–115.
- Carnevale, P. J., & Pruitt, D. G. (1992). Negotiation and mediation. *Annual Review of Psychology*, 43(1), 531–582.
- Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (RoSAS): development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 254–262).
- Cartwright, D. E. (1988). Schopenhauer's compassion and Nietzsche's



- pity. *Schopenhauer Jahrbuch*, 69, 557–567.
- Castiello, U. (2003). Understanding other people's actions: Intention and attention. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), 416.
- Cheon, E., Sher, S. T.-H., Sabanović, Š., & Su, N. M. (2019). I beg to differ: Soft conflicts in collaborative design using design fictions. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 201–214). New York, NY, USA: Association for Computing Machinery.
- Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, 66(9), 1261–1266.
- Chugh, D., Bazerman, M. H., & Banaji, M. R. (2005). Bounded ethicality as a psychological barrier to recognizing conflicts of interest. In D. C. D. Moore G. Loewenstein & M. H. Bazerman (Eds.), *Conflicts of Interest: Challenges and Solutions in Business, Law, Medicine, and Public Policy* (pp. 74–95). New York, NY, USA: Cambridge University Press.
- Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, 58(1), 7–19.
- Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., ... Cowan, B. (2020). The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers*.
- Coeckelbergh, M. (2009). Virtual moral agency, virtual moral responsibility: on the moral significance of the appearance, perception, and performance of artificial agents. *AI & Society*, 24(2), 181–189.
- Cosmides, L., & Tooby, J. (2008). Can a general deontic logic capture the facts of human moral reasoning? How the mind interprets social exchange rules and detects cheaters. In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral Psychology: The Evolution of Morality: Adaptations and Innateness* (pp. 114–137). Cambridge, MA, USA.
- Cowie, R. (2015). Ethical issues in affective computing. In D. S. G. J. Calvo R. A. & A. Kappas (Eds.), *The Oxford Handbook of Affective Computing* (p. 334–338). New York, NY, USA: Oxford University Press.
- Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., ... Whittaker, M. (2019). *AI Now Report 2019*. AI Now Institute at New York University.
- Creswell, J., Clark, V., Gutmann, M., & Hanson, W. (2008). An expanded typology for classifying mixed methods research into designs. *The Mixed Methods Reader*, 159–96.
- Creswell, J. W., & Creswell, J. D. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

- Cuddy, A. J., Fiske, S. T., Kwan, V. S., Glick, P., Demoulin, S., Leyens, J.-P., ... others (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1), 1–33.
- Cuijpers, R. H., & Knops, M. A. M. H. (2015). Motions of robots matter! The social effects of idle and meaningful motions. In A. Tapus, E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social robotics* (pp. 174–183). Cham: Springer International Publishing.
- Curry, O. S., Chesters, M. J., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of ‘morality-as-cooperation’ with a new questionnaire. *Journal of Research in Personality*, 78, 106–124.
- Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford Handbook of Moral Psychology*, 47–71.
- Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of oz studies: Why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces* (p. 193–200). New York, NY, USA: Association for Computing Machinery.
- Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, 22(5), 811–817.
- Damasio, A. R. (2006). *Descartes’ Error: Emotion, Reason, and the Human Brain*. New York, NY, USA: Random House.
- Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology*, 18(4), 299–309.
- Danaher, J. (2019). Welcoming robots into the moral circle: A defence of ethical behaviourism. *Science and Engineering Ethics*, 1–27.
- Danielson, P. (2009). Can robots have a conscience? *Nature*, 457(7229), 540.
- Darley, J. M., & Batson, C. D. (1973). From Jerusalem to Jericho: A study of situational and dispositional variables in helping behavior. *Journal of Personality and Social Psychology*, 27(1), 100.
- Darwall, S. (2004). Respect and the second-person standpoint. In *Proceedings and Addresses of the American Philosophical Association* (Vol. 78, pp. 43–59).
- Darwall, S. (2006). *The Second-person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA, USA: Harvard University Press.
- Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 362(1480), 679–704.
- Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, 22(1), 209–220.
- Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. R. (1994). Facili-

- tating internalization: The self-determination theory perspective. *Journal of Personality*, 62(1), 119–142.
- De Graaf, M. M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*. Palo Alto, CA, USA: AAAI Press.
- de Melo, C. M., Carnevale, P. J., Read, S. J., & Gratch, J. (2014). Reading people’s minds from emotion expressions in interdependent decision making. *Journal of Personality and Social Psychology*, 106(1), 73–88.
- de Melo, C. M., & Gratch, J. (2015, sep). People show envy, not guilt, when making decisions with machines. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (p. 315–321). Los Alamitos, CA, USA: IEEE Computer Society.
- de Melo, C. M., Gratch, J., & Carnevale, P. J. (2014). The importance of cognition and affect for artificially intelligent decision makers. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*. Palo Alto, CA, USA: AAAI Press.
- de Melo, C. M., Gratch, J., & Carnevale, P. J. (2015). Humans versus computers: Impact of emotion expressions on people’s decision making. *IEEE Transactions on Affective Computing*, 6(2), 127–136.
- de Melo, C. M., Khooshabeh, P., Amir, O., & Gratch, J. (2018). Shaping cooperation between humans and agents with emotion expressions and framing. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2018)* (pp. 2224–2226).
- Dennett, D. (1989). *The Intentional Stance*. Cambridge, MA, USA: MIT press.
- Dennett, D. (2008). *Kinds of Minds: Toward an Understanding of Consciousness*. New York, NY, USA: Basic Books.
- Dennett, D. (2009). Darwin’s “strange inversion of reasoning”. *Proceedings of the National Academy of Sciences*, 106(Supplement 1), 10061–10065.
- Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, 48(4), 87–106.
- Dennett, D. C. (2017). *From Bacteria to Bach and Back: The Evolution of Minds*. New York, NY, USA: WW Norton & Company.
- de Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012). The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, pp. 263–267).
- de Weerd, H., Verbrugge, R., & Verheij, B. (2017). Negotiating with other minds: The role of recursive theory of mind in negotiation with incomplete information. *Autonomous Agents and Multi-Agent*

- Systems*, 31(2), 250–287.
- Dodge, R., Daly, A. P., Huyton, J., & Sanders, L. D. (2012). The challenge of defining wellbeing. *International Journal of Wellbeing*, 2(3), 222–235.
- Doris, J. M. (1998). Persons, situations, and virtue ethics. *Nous*, 32(4), 504–530.
- Dotsch, R., & Wigboldus, D. H. (2008). Virtual prejudice. *Journal of Experimental Social Psychology*, 44(4), 1194–1198.
- Duff, R. (2003). Probation, punishment and restorative justice: Should al turism be engaged in punishment? *The Howard Journal of Criminal Justice*, 42(2), 181–197.
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3), 177–190.
- Dunne, A., & Raby, F. (2013). *Speculative everything: Design, fiction, and social dreaming*. Cambridge, MA, USA: MIT press.
- Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28), 7900–7905.
- Ekman, P. (1992). Are there basic emotions? *Psychological Review*, 99(3), 550–553.
- Ekman, P. (1993). Facial expression and emotion. *American Psychologist*, 48(4), 384–392.
- Ekman, P., Friesen, W. V., & Ellsworth, P. (2013 [1972]). *Emotion in the human face: Guidelines for research and an integration of findings* (Vol. 11). Elmsford, NY, USA: Pergamon Press Inc.
- Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, 14(4), 583–610.
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886.
- Eyssel, F., Kuchenbrandt, D., & Bobinger, S. (2011). Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 61–68).
- Falconer, C. J., King, J. A., & Brewin, C. R. (2015). Demonstrating mood repair with a situation-based measure of self-compassion and self-criticism. *Psychology and Psychotherapy: Theory, Research and Practice*, 88(4), 351–365.
- Falconer, C. J., Rovira, A., King, J. A., Gilbert, P., Antley, A., Fearon, P., ... Brewin, C. R. (2016). Embodying self-compassion within virtual reality and its effects on patients with depression. *BJPsych Open*, 2(1), 74–80.
- Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a

- prototype perspective. *Journal of Experimental Psychology: General*, 113(3), 464–486.
- Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, & M.-A. Williams (Eds.), *Social Robotics* (pp. 199–208). Heidelberg, Germany: Springer.
- Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6), 878–902.
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health*, 4(2).
- Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines*, 18(3), 303–329.
- Floridi, L. (2013). *The Ethics of Information*. New York, NY, USA: Oxford University Press.
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.
- Fogg, B., & Nass, C. (1997). How users reciprocate to computers: an experiment that demonstrates behavior change. In *CHI'97 Extended Abstracts on Human Factors in Computing Systems* (pp. 331–332). New York, NY, USA: Association for Computing Machinery.
- Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 5–15.
- Forssell, R. (2016). Exploring cyberbullying and face-to-face bullying in working life—prevalence, targets and expressions. *Computers in Human Behavior*, 58, 454–460.
- Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3), 347–369.
- Foucault, M. (2006 [1961]). *History of Madness*. Abingdon, UK: Routledge.
- Foucault, M. (2012 [1975]). *Discipline and Punish: The Birth of the Prison*. New York, NY, USA: Vintage.
- Frank, L. E. (2019). What Do We Have to Lose? Offloading Through Moral Technologies: Moral Struggle and Progress. *Science and Engineering Ethics*, 1–17.
- Fricker, M. (2016). What's the point of blame? A paradigm based explanation. *Noûs*, 50(1), 165–183.
- Friedman, B. (1996). Value-sensitive design. *Interactions*, 3(6), 16–23.

- Friedman, B., Kahn Jr., P. H., & Borning, A. (2008). Value sensitive design and information systems. In *The Handbook of Information and Computer Ethics* (p. 69-101). Hoboken, NJ, USA: John Wiley Sons, Ltd.
- Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, 43(5), 349-358.
- Frijda, N. H., Manstead, A. S., & Bem, S. (2000). *Emotions and Beliefs: How Feelings Influence Thoughts*. Cambridge, UK: Cambridge University Press.
- Furlough, C., Stokes, T., & Gillan, D. J. (2019). Attributing blame to robots: I. the influence of robot autonomy. *Human Factors*, 1-11.
- Gaver, W. W., Beaver, J., & Benford, S. (2003). Ambiguity as a resource for design. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 233-240). New York, NY, USA: Association for Computing Machinery.
- Gendron, M., & Feldman Barrett, L. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion Review*, 1(4), 316-339.
- Gergen, K. J. (1995). Metaphor and monophony in the 20th-century psychology of emotions. *History of the Human Sciences*, 8(2), 1-23.
- Germer, C. K., & Neff, K. D. (2013). Self-compassion in clinical practice. *Journal of Clinical Psychology*, 69(8), 856-867.
- Gilbert, P. (2014). The origins and nature of compassion focused therapy. *British Journal of Clinical Psychology*, 53(1), 6-41.
- Gino, F., & Mogilner, C. (2014). Time, money, and morality. *Psychological Science*, 25(2), 414-421.
- Giorgi, A. (2012). The descriptive phenomenological psychological method. *Journal of Phenomenological Psychology*, 43(1), 3-12.
- Godman, M., & Jefferson, A. (2017). On blaming and punishing psychopaths. *Criminal Law and Philosophy*, 11(1), 127-142.
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*. New York, NY, USA: Doubleday.
- Goldstein, N. J., & Cialdini, R. B. (2007). The spyglass self: A model of vicarious self-perception. *Journal of Personality and Social Psychology*, 92(3), 402-417.
- Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, 106(1), 148-168.
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In P. Devine & A. Plant (Eds.), *Advances in Experimental Social Psychology* (Vol. 47, pp. 55-130). London, UK: Elsevier.
- Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H.

- (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, 101(2), 366–385.
- Gratch, J., DeVault, D., Lucas, G. M., & Marsella, S. (2015). Negotiation as a challenge problem for virtual humans. In W.-P. Brinkman, J. Broekens, & D. Heylen (Eds.), *Intelligent Virtual Agents* (pp. 201–215). Cham: Springer International Publishing.
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619–619.
- Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology*, 3(3), 405–423.
- Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, 143(4), 1600–1615.
- Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, 23(2), 206–215.
- Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, 96(3), 505–520.
- Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, 125(1), 125–130.
- Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23(2), 101–124.
- Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12), 517–523.
- Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, 11(8), 322–323.
- Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400.
- Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fmri investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105–2108.
- Guéguen, N. (2012). The sweet smell of... implicit helping: Effects of pleasant ambient fragrance on spontaneous help in shopping malls. *The Journal of Social Psychology*, 152(4), 397–400.
- Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4), 367–388.

- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4), 814–834.
- Haidt, J. (2003). The moral emotions. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of Affective Sciences* (Vol. 11, pp. 852–870). Oxford, UK: Oxford University Press.
- Haidt, J. (2008). Morality. *Perspectives on Psychological Science*, 3(1), 65–72.
- Han, B.-C. (2017). *Psychopolitics. Neoliberalism and New Technologies of Power*. London, UK: Verso.
- Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (p. 929–932). New York, NY, USA: Association for Computing Machinery.
- Haslam, N. (2012). Morality, mind, and humanness. *Psychological Inquiry*, 23(2), 172–174.
- Haugeland, J. (1989). *Artificial Intelligence: The Very Idea*. Cambridge, MA, USA: MIT Press.
- Hellström, T., & Bensch, S. (2018). Understandable robots: What, why, and how. *Paladyn, Journal of Behavioral Robotics*, 9(1), 110–123.
- Hess, U., & Fischer, A. (2013). Emotional mimicry as social regulation. *Personality and Social Psychology Review*, 17(2), 142–157.
- Hewig, J., Trippe, R. H., Hecht, H., Coles, M. G., Holroyd, C. B., & Miltner, W. H. (2008). An electrophysiological analysis of coaching in blackjack. *Cortex*, 44(9), 1197–1205.
- High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from High-Level Expert Group on Artificial Intelligence (AI HLEG) of the European Commission. (Available at <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top>)
- Howells, A., Ivtzan, I., & Eiroa-Orosa, F. J. (2016). Putting the ‘app’ in happiness: A randomised controlled trial of a smartphone-based mindfulness intervention to enhance wellbeing. *Journal of Happiness Studies*, 17(1), 163–185.
- Hsu, T. (2010, November). *Japanese pop star hatsune miku takes the stage – as a 3-d hologram*. <https://latimesblogs.latimes.com/technology/2010/11/japanese-pop-star-takes-the-stage-as-a-3-d-hologram.html>. (Accessed on 01/29/2020)
- Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences*, 13(1), 1–6.
- Hume, D. (2003 [1739]). *A Treatise of Human Nature*. Mineola, NY, USA: Dover Publications.
- Hutcheson, F. (2008 [1725]). *An Inquiry into the Original of Our Ideas of Beauty and Virtue: In Two Treatises*. Indianapolis, IN, USA: Liberty



Fund.

- Isen, A. M., & Levin, P. F. (1972). Effect of feeling good on helping: cookies and kindness. *Journal of Personality and Social Psychology*, 21(3), 384–388.
- Jameson, F. (1991). *Postmodernism, or, the Cultural Logic of Late Capitalism*. Durham, NC, USA: Duke University Press.
- Jeffrey, C. (2018, November). *Japanese man marries anime hologram of Hatsune Miku - TechSpot*. <https://www.techspot.com/news/77385-japanese-man-marries-anime-hologram-hatsune-miku.html>. (Accessed on 01/29/2020)
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71.
- Johnson, D. O., & Cuijpers, R. H. (2019). Investigating the effect of a humanoid robot's head position on imitating human emotions. *International Journal of Social Robotics*, 11(1), 65–74.
- Kahneman, D., & Riis, J. (2005). Living, and thinking about it: Two perspectives on life. In B. N. Huppert F.A. & B. Keverne (Eds.), *The Science of Well-being* (pp. 285–304). Oxford, UK: Oxford University Press.
- Kant, I. (1964 [1785]). *Groundwork of the Metaphysic of Morals*. New York, NY, USA: Harper and Row Publishers. (Translated by H. J. Patton)
- Kant, I. (1996 [1797]). *The Metaphysics of Morals*. Cambridge, UK: Cambridge University Press. (Translated by Mary J. Gregor)
- Kant, I. (1998 [1781]). *Critique of Pure Reason*. Cambridge, UK: Cambridge University Press. (Translated by P. Guyer and A.W. Wood)
- Karreman, D. E., Ludden, G. D., & Evers, V. (2019). Beyond R2D2: Designing Multimodal Interaction Behavior for Robot-specific Morphology. *ACM Transactions on Human-Robot Interaction (THRI)*, 8(3), 1–32.
- Kawachi, I., & Berkman, L. F. (2001). Social ties and mental health. *Journal of Urban Health*, 78(3), 458–467.
- Keijzers, M., & Bartneck, C. (2018). Mindless robots get bullied. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (p. 205–214). New York, NY, USA: Association for Computing Machinery.
- Keltner, D. (1995). Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *Journal of Personality and Social Psychology*, 68(3), 441–454.
- Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral type-casting accounts. *Cognition*, 146, 33–47.

- Kierkegaard, S. (2000 [1835-1855]). *The Essential Kierkegaard* (H. Hong & E. Hong, Eds.). Princeton, NJ, USA: Princeton University Press.
- Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 80–85).
- Kitzinger, J. (1995). Qualitative research: Introducing focus groups. *BMJ*, 311(7000), 299–302.
- Kizilcec, R. F. (2016). How much information? effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (p. 2390–2395). New York, NY, USA: Association for Computing Machinery.
- Klimecki, O. M., Leiberg, S., Ricard, M., & Singer, T. (2013). Differential pattern of functional brain plasticity after compassion and empathy training. *Social Cognitive and Affective Neuroscience*, 9(6), 873–879.
- Klincewicz, M. (2017). Challenges to engineering moral reasoners: Time and context. In R. J. Patrick Lin Keith Abney (Ed.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence* (pp. 244–257). New York, NY, USA: Oxford University Press.
- Klincewicz, M. (2019). Robotic nudges for moral improvement through stoic practice. *Techné: Research in Philosophy and Technology*, 23(3), 425–455.
- Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63(3), 190–194.
- Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908–911.
- Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. Goslin (Ed.), (pp. 347–480). Chicago, IL, USA: Rand McNally.
- Kohlberg, L. (1971). Stages of moral development. *Moral Education*, 1(51), 23–92.
- Kohlberg, L. (1973). "Continuity in Childhood and Adult Moral Development." *Collected Papers on Moral Development and Moral Education*. Cambridge, MA, USA: Harvard University Press.
- Kohlberg, L. (1984). *Essays on Moral Development: The Psychology of Moral Development* (Vol. 2). San Francisco, CA, USA: Harper & Row.
- Kolata, G. (2019, October). *You Got a Brain Scan at the Hospital. Someday a Computer May Use It to Identify You* - *The New York Times*. <https://www.nytimes.com/2019/10/23/health/brain-scans-personal-identity.html>. (Accessed on 01/29/2020)
- Komatsu, T. (2016). How do people judge moral wrongness in a robot

- and in its designers and owners regarding the consequences of the robot's behaviors? In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 1168–1171).
- Kongtrul, J. (1987). *The Great Path of Awakening: The Classic Guide to Lojong, a Tibetan Buddhist Practice for Cultivating the Heart of Compassion*. Boston, MA, USA: Shambhala Publications. (Translated by Ken McLeod)
- Konrath, S. H., O'Brien, E. H., & Hsing, C. (2011). Changes in dispositional empathy in American college students over time: A meta-analysis. *Personality and Social Psychology Review*, 15(2), 180–198.
- Kontogiorgos, D., Skantze, G., Abelho Pereira, A. T., & Gustafson, J. (2019). The effects of embodiment and social eye-gaze in conversational agents. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society (CogSci)* (pp. 589–595).
- Krämer, N. C. (2008). Theory of mind as a theoretical prerequisite to model communication with virtual humans. In I. Wachsmuth & G. Knoblich (Eds.), *Modeling Communication with Robots and Virtual Humans* (pp. 222–240). Heidelberg, Germany: Springer.
- Krämer, N. C., von der Pütten, A., & Eimler, S. (2012). Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In M. Zacarias & J. V. de Oliveira (Eds.), *Human-Computer Interaction: The Agency Perspective*. Springer, address=.
- Kraus, M. W., Onyeador, I. N., Daumeyer, N. M., Rucker, J. M., & Richeson, J. A. (2019). The misperception of racial economic inequality. *Perspectives on Psychological Science*, 14(6), 899–921.
- Krebs, D. L., & Denton, K. (2005). Toward a more pragmatic approach to morality: a critical evaluation of Kohlberg's model. *Psychological Review*, 112(3), 629–649.
- Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, 32(9), 509–515.
- Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613.
- Krueger, R. A. (2014). *Focus groups: A Practical Guide for Applied Research*. 5th Edition. Thousand Oaks, CA, USA: Sage publications.
- Lapsley, D. K. (1996). *Moral Psychology*. New York, NY, USA: Westview Press.
- Lazarus, R. S. (2006). Emotions and interpersonal relationships: Toward a person-centered conceptualization of emotions and coping. *Journal of Personality*, 74(1), 9–46.
- Leahu, L., & Sengers, P. (2014). Freaky: performing hybrid human-

- machine emotion. In *Proceedings of the 2014 Conference on Designing Interactive Systems* (pp. 607–616). New York, NY, USA: Association for Computing Machinery.
- Leary, M. R., Tate, E. B., Adams, C. E., Batts Allen, A., & Hancock, J. (2007). Self-compassion and reactions to unpleasant self-relevant events: The implications of treating oneself kindly. *Journal of Personality and Social Psychology*, 92(5), 887–904.
- Lee, M. (2020). Speech acts redux: Beyond request-response interactions. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. New York, NY, USA: Association for Computing Machinery.
- Lee, M., Ackermans, S., van As, N., Chang, H., Lucas, E., & IJsselsteijn, W. (2019). Caring for Vincent: A chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery.
- Lee, M., Frank, L., Beute, F., de Kort, Y., & IJsselsteijn, W. (2017). Bots mind the social-technical gap. In *Proceedings of 15th European Conference on Computer-Supported Cooperative Work-Exploratory Papers*.
- Lee, M., Lucas, G., Mell, J., Johnson, E., & Gratch, J. (2019). What's on your virtual mind? Mind perception in human-agent negotiations. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 38–45).
- Lei, S., & Gratch, J. (2019). Smiles signal surprise in a social dilemma. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 627–633). IEEE Computer Society.
- Lemaignan, S., Fink, J., Mondada, F., & Dillenbourg, P. (2015). You're doing it wrong! Studying unexpected behaviors in child-robot interaction. In A. Tapus, E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social robotics* (pp. 390–400). Cham: Springer International Publishing.
- Levin, D. T., Killingsworth, S. S., Saylor, M. M., Gordon, S. M., & Kawamura, K. (2013). Tests of concepts about different kinds of minds: Predictions about the behavior of computers, robots, and people. *Human-Computer Interaction*, 28(2), 161–191.
- Lewin, K. (1931). The conflict between Aristotelian and Galileian modes of thought in contemporary psychology. *The Journal of General Psychology*, 5(2), 141–177.
- Lewin, K. (1951). *Field Theory in Social Science: Selected Theoretical Papers* (D. Cartwright, Ed.). New York, NY, USA: Harpers and Brothers.
- Lim, A., & Okuno, H. G. (2015). A recipe for empathy. *International Journal of Social Robotics*, 7(1), 35–49.
- Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not ex-

- planations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2119–2128). New York, NY, USA: Association for Computing Machinery.
- Lindley, J., & Coulton, P. (2015). Back to the future: 10 years of design fiction. In *Proceedings of the 2015 British HCI Conference* (pp. 210–211).
- Lomas, M., Chevalier, R., Cross, E. V., Garrett, R. C., Hoare, J., & Kopack, M. (2012). Explaining robot actions. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (p. 187–188). New York, NY, USA: Association for Computing Machinery.
- Lottridge, D., Chignell, M., & Jovicic, A. (2011). Affective interaction: Understanding, evaluating, and designing for human emotion. *Reviews of Human Factors and Ergonomics*, 7(1), 197–217.
- Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37, 94–100.
- Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., ... Shively, R. (2017). Shaping trust through transparent design: Theoretical and experimental guidelines. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in Human Factors in Robots and Unmanned Systems* (pp. 127–136). Cham: Springer International Publishing.
- Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? people apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). New York, NY, USA: Association for Computing Machinery.
- Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. In T. B. Klaus R. Scherer & E. Roesch (Eds.), *A Blueprint for Affective Computing: A Sourcebook and Manual* (Vol. 11, pp. 21–46). New York, NY, USA: Oxford University Press.
- Marsella, S. C., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1), 70–90.
- Mascolo, M. F. (2016). Beyond objectivity and subjectivity: The intersubjective foundations of psychological science. *Integrative Psychological and Behavioral Science*, 50(4), 543–554.
- Mateas, M., & Sengers, P. (1999). Narrative intelligence. In *Narrative Intelligence: Papers from the AAAI Fall Symposium (1999)*, AAAI TR FS-99-01.
- Matsuzoe, S., & Tanaka, F. (2012). How smartly should robots behave?: Comparative investigation on the learning ability of a

- care-receiving robot. In *Proceedings of the 21th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN '12)* (pp. 339–344). IEEE.
- Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183.
- McArthur, N., & Twist, M. L. (2017). The rise of digisexuality: Therapeutic challenges and possibilities. *Sexual and Relationship Therapy*, 32(3-4), 334–344.
- McCarthy, E. D. (1994). The social construction of emotions: New directions from culture theory. *Social Perspectives on Emotion*, 2(1), 267–279.
- McDermott, D. (2001). The permissibility of punishment. *Law and Philosophy*, 20(4), 403–432.
- McFatter, R. M. (1978). Sentencing strategies and justice: Effects of punishment philosophy on sentencing decisions. *Journal of Personality and Social Psychology*, 36(12), 1490–1500.
- McLaughlin, D., & Brody, B. (2019, December). *FTC Eyes Suit to Block Facebook Plan to Merge App Systems - Bloomberg*. <https://www.bloomberg.com/news/articles/2019-12-12/u-s-ftc-eyes-suit-to-block-facebook-plan-to-merge-app-systems>. (Accessed on 01/29/2020)
- McRae, E. (2012). The psychology of moral judgment and perception in Indo-Tibetan Buddhist ethics. In D. Cozort & J. Shields (Eds.), *The Oxford Handbook of Buddhist Ethics* (pp. 335–358). Oxford, UK: Oxford University Press.
- McRae, E. (2018). Finding a place for Buddhism in the ethics of the future: Comments on Shannon Vallor's *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. *Philosophy and Technology*, 31, 277–282.
- McStay, A. (2018). *Emotional AI: The Rise of Empathic Media*. London, UK: Sage.
- Mell, J., & Gratch, J. (2017). Grumpy & Pinocchio: Answering human-agent negotiation questions through realistic agent design. In *Proceedings of the 16th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2017)* (p. 401–409). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.
- Mell, J., Lucas, G., & Gratch, J. (2015). An effective conversation tactic for creating value over repeated negotiations. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)* (p. 1567–1576). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

- Mell, J., Lucas, G., & Gratch, J. (2017). Prestige questions, online agents, and gender-driven differences in disclosure. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, & S. Kopp (Eds.), *Proceedings of the 17th International Conference on Intelligent Virtual Agents* (pp. 273–282). Cham: Springer International Publishing.
- Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for Multi-UxV management. *Human Factors*, 58(3), 401–415.
- Midgley, M. (1996). *Utopias, Dolphins, and Computers: Problems of Philosophical Plumbing*. London, UK: Routledge.
- Misselhorn, C. (2015). Collective agency and cooperation in natural and artificial systems. In C. Misselhorn (Ed.), *Collective Agency and Cooperation in Natural and Artificial Systems* (pp. 3–24). Springer.
- Monin, B., Pizarro, D. A., & Beer, J. S. (2007). Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology*, 11(2), 99–111.
- Moon, Y., & Nass, C. (1996). How “real” are computer personalities? psychological responses to personality types in human-computer interaction. *Communication Research*, 23(6), 651–674.
- Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.
- Moore, R. C., Martin, A. S., Kaup, A. R., Thompson, W. K., Peters, M. E., Jeste, D. V., ... Eyler, L. T. (2015). From suffering to caring: A model of differences among older adults in levels of compassion. *International Journal of Geriatric Psychiatry*, 30(2), 185–191.
- Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33–35.
- Morris, M. W., & Keltner, D. (1999). How emotions work: An analysis of the social functions of emotional expression in negotiation. In B. Staw & R. Sutton (Eds.), *Research in Organizational Behavior* (Vol. 11, pp. 1–50). Amsterdam, the Netherlands: JAI.
- Moshman, D. (2004). False moral identity: Self-serving denial in the maintenance of moral self-conceptions. In D. K. Lapsley & D. Narvaez (Eds.), *Moral Development, Self, and Identity* (Vol. 2, pp. 83–109). Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.
- Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5-6), 527–539.
- Muller, M., & Erickson, T. (2018). In the data kitchen: A review (a design fiction on data science). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–10). New York, NY, USA: Association for Computing Machinery.

- Myers, D. G., & Diener, E. (1995). Who is happy? *Psychological Science*, 6(1), 10–19.
- Müller, V. C. (2020). Ethics of Artificial Intelligence and Robotics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2020/entries/ethics-ai/>.
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 72–78). New York, NY, USA: Association for Computing Machinery.
- Neff, K. (2003a). Self-compassion: An alternative conceptualization of a healthy attitude toward oneself. *Self and Identity*, 2(2), 85–101.
- Neff, K. (2003b). *Test how self-compassionate you are*. <https://self-compassion.org/test-how-self-compassionate-you-are/>. (Accessed: 2018-07-01)
- Neff, K. (2008). *Self-compassion exercises*. <http://self-compassion.org/category/exercises/#exercises>. (Accessed: 2018-07-01)
- Neff, K. (2018). *Why women need fierce self-compassion*. [https://greatergood.berkeley.edu/article/item/why\\_women\\_need\\_fierce\\_self\\_compassion](https://greatergood.berkeley.edu/article/item/why_women_need_fierce_self_compassion). Greater Good Magazine. (Accessed: 2018-12-30)
- Nehaniv, C. L. (1999). Narrative for artifacts: Transcending context and self. In *Narrative Intelligence: Papers from the AAAI Fall Symposium* (1999), AAAI TR FS-99-01.
- Nevada, Department of Motor Vehicles. (2012). *Nevada DMV Issues First Autonomous Vehicle Testing License to Google*. <https://dmv.nv.com/news/12005-autonomous-vehicle-licensed.html>. Nevada Department of Motor Vehicles.
- Nichols, S. (2004). *Sentimental Rules: On the Natural Foundations of Moral Judgment*. New York, NY, USA: Oxford University Press.
- Nietzsche, F. (1989 [1885]). *Beyond Good and Evil : Prelude to a Philosophy of the Future*. New York, NY, USA: Vintage Books.
- Nimako, K., Abdou, A., & Willemsen, G. (2014). Chattel slavery and racism: A reflection on the Dutch experience. In P. Essed & I. Hoving (Eds.), *Dutch Racism* (pp. 31–51). Amsterdam, the Netherlands: Brill Rodopi.
- Noddings, N. (2008). Caring and moral education. In L. P. Nucci & D. Narvaez (Eds.), *Handbook of Moral and Character Education* (pp. 161–174). New York, NY, USA: Routledge.
- Nucci, L. (2004). Reflections on the moral self construct. In D. K. Lapsley & D. Narvaez (Eds.), *Moral Development, Self, and Identity* (Vol. 2, pp. 111–132). Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.



- Nussbaum, M. (1996). Compassion: The basic social emotion. *Social Philosophy and Policy*, 13(1), 27–58.
- Nussbaum, M. C. (2001). *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge, UK: Cambridge University Press.
- Nutt, A. E. (2017). “The Woebot will see you now” - the rise of chatbot therapy. <https://www.washingtonpost.com/news/to-your-health/wp/2017/12/03/the-woebot-will-see-you-now-the-rise-of-chatbot-therapy/>. The Washington Post. (Accessed: 2018-07-16)
- Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, 24(4), 1201–1219.
- Nyholm, S. (2020). *Humans and Robots: Ethics, Agency, and Anthropomorphism*. London, UK: Rowman & Littlefield Publishers.
- Nyholm, S., & Frank, L. E. (2017). From sex robots to love robots: Is mutual love with a robot possible? In J. Danaher & N. McArthur (Eds.), *Robot Sex: Social and Ethical Implications* (pp. 219–243). Cambridge, MA, USA: MIT press.
- Nyholm, S., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: An applied trolley problem? *Ethical Theory and Moral Practice*, 19(5), 1275–1289.
- Ohtsubo, Y. (2007). Perceived intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect. *Japanese Psychological Research*, 49(2), 100–110.
- Olson, P. (2016, 2). *Get ready for the chat bot revolution: They’re simple, cheap and about to be everywhere - forbes*. <https://www.forbes.com/sites/parmyolson/2016/02/23/chat-bots-facebook-telegram-wechat/#7ce3f8802068>. Forbes. (Accessed on 12/16/2019)
- Oosterbeek, H., Sloof, R., & Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2), 171–188.
- Osofsky, M. J., Bandura, A., & Zimbardo, P. G. (2005). The role of moral disengagement in the execution process. *Law and Human Behavior*, 29(4), 371–393.
- Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Services Research*, 34(5 Pt 2), 1189–1208.
- Piazza, J., Landy, J. F., & Goodwin, G. P. (2014). Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition*, 131(1), 108–124.
- Picard, R. W. (1995). Affective computing. Technical Report 321, MIT Media Lab, Perceptual Computing Group.
- Picard, R. W. (2003). Affective computing: Challenges. *International*

- Journal of Human-Computer Studies*, 59(1-2), 55–64.
- Plato. (2002 [±400-348 BC]). *Plato: Five Dialogues: Euthyphro, Apology, Crito, Meno, Phaedo. Second Edition* (J. M. Cooper, Ed.). Indianapolis, IN, USA: Hackett Publishing. (Translated by G. M. A. Grube)
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4), 515–526.
- Prinz, J. (2004). *Gut Reactions: A Perceptual Theory of Emotion*. New York, NY, USA: Oxford University Press.
- Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations*, 9(1), 29–43.
- Prinz, J. (2008). Is morality innate? In W. Sinnott-Armstrong (Ed.), *Moral psychology. The Evolution of Morality: Adaptations and Innateness* (Vol. 1, pp. 367–406). Cambridge, MA, USA: MIT Press.
- Pruitt, D. G. (1967). Reward structure and cooperation: The decomposed prisoner's dilemma game. *Journal of Personality and Social Psychology*, 7(1, part 1), 21–27.
- Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology*, 28(1), 363–392.
- Puka, B. (2004). Altruism and character. In D. K. Lapsley & D. Narvaez (Eds.), *Moral Development, Self, and Identity* (Vol. 2, pp. 161–188). Mahwah, NJ, USA: Lawrence Erlbaum Associates Publishers.
- Reeves, B., & Nass, C. I. (1996). *The Media Equation: How People Treat computers, Television, and New Media like Real People and Places*. New York, NY, USA: Cambridge University Press.
- Reeves, G. (2012). A perspective on ethics in the Lotus Sūtra. In D. Cozort & J. Shields (Eds.), *The Oxford Handbook of Buddhist Ethics* (pp. 335–358). Oxford University Press.
- Rhue, L. (2018). Racial influence on automated perceptions of emotions. Available at SSRN: <https://ssrn.com/abstract=3281765>.
- Richardson, T., Wrightman, M., Yeebo, M., & Lisicka, A. (2017). Reliability and score ranges of the PHQ-9 and GAD-7 in a primary and secondary care mental health service. *Journal of Psychosocial Rehabilitation and Mental Health*, 4(2), 237–240.
- Rodrigues, D., Prada, M., Gaspar, R., Garrido, M. V., & Lopes, D. (2018). Lisbon emoji and emoticon database (LEED): Norms for emoji and emoticons in seven evaluative dimensions. *Behavior Research Methods*, 50(1), 392–405.
- Rossmly, B., Völkel, S. T., Naphausen, E., Kimm, P., Wiethoff, A., & Muxel, A. (2020). Punishable AI: Examining users' attitude towards robot punishment. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (p. 179–191). New York, NY, USA: Association for Computing Machinery.
- Ruijten, P. A. (2015). *Responses to Human-like Artificial Agents*. PhD

- thesis, Eindhoven University of Technology. Hertogenbosch, The Netherlands: Uitgeverij BOXPress.
- Ruijten, P. A., Bouten, D. H. L., Rouschop, D. C. J., Ham, J., & Midden, C. J. H. (2014). Introducing a Rasch-type anthropomorphism scale. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 280–281). ACM.
- Ruijten, P. A., & Cuijpers, R. H. (2017). Dynamic perceptions of human-likeness while interacting with a social robot. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 273–274). ACM.
- Ruijten, P. A., Midden, C. J., & Ham, J. (2015). Lonely and susceptible: The influence of social exclusion and gender on persuasion by an artificial agent. *International Journal of Human-Computer Interaction*, 31(11), 832–842.
- Russell, J. A., Bachorowski, J.-A., & Fernández-Dols, J.-M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, 54(1), 329–349.
- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76(5), 805–819.
- Šabanović, S. (2010). Robots in society, society in robots. *International Journal of Social Robotics*, 2(4), 439–450.
- Salem, M., Eyssel, F., Rohlfing, K., Kopp, S., & Joublin, F. (2011). Effects of gesture on the perception of psychological anthropomorphism: A case study with a humanoid robot. In B. Mutlu, C. Bartneck, J. Ham, V. Evers, & T. Kanda (Eds.), *Social Robotics* (pp. 31–41). Heidelberg, Germany: Springer.
- Salvini, P., Ciaravella, G., Yu, W., Ferri, G., Manzi, A., Mazzolai, B., ... Dario, P. (2010). How safe are service robots in urban environments? bullying a robot. In *19th International Symposium in Robot and Human Interactive Communication (ROMAN)* (p. 1-7).
- Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626), 1755–1758.
- Sartre, J.-P. (2016). *What is Subjectivity?* London, UK: Verso Books.
- Scanlon, T. M. (2013). Interpreting blame. In N. A. T. D. Justin Coates (Ed.), *Blame. Its Nature and Norms* (pp. 84–99).
- Scarantino, A. (2017). How to do things with emotional expressions: The theory of affective pragmatics. *Psychological Inquiry*, 28(2-3), 165–185.
- Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Höllerer, T. (2019). I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User*

- Interfaces* (p. 240–251). New York, NY, USA: Association for Computing Machinery.
- Scherer, K. R., Bänziger, T., & Roesch, E. (Eds.). (2010). *A Blueprint for Affective Computing: A Sourcebook and Manual*. Oxford, UK: Oxford University Press.
- Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, 1096–1109.
- Schopenhauer, A. (1995 [1840]). *On the Basis of Morality*. Indianapolis, IN, USA: Hackett Publishing.
- Schulte, B. F., Marshall, P., & Cox, A. L. (2016). Homes for life: A design fiction probe. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*. New York, NY, USA: Association for Computing Machinery.
- Sengers, P. (2000). Narrative intelligence. In K. Dautenhahn (Ed.), *Human Cognition and Social Agent Technology* (Vol. 19, pp. 1–26). Amsterdam, the Netherlands: John Benjamins Publishing.
- Shalvi, S., Dana, J., Handgraaf, M. J., & De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2), 181–190.
- Shantideva, A. (1979). *A Guide to the Bodhisattva's Way of Life*. Dharamsala, India, Library of Tibetan Works and Archives. (Translated by Stephen Batchelor)
- Shoemaker, D. (2013). Blame and punishment. In N. A. T. D. Justin Coates (Ed.), *Blame. Its Nature and Norms* (pp. 100–118). Oxford, UK: Oxford University Press.
- Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! An interaction with a cheating robot. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (p. 219–226). IEEE Press.
- Siegel, M., Breazeal, C., & Norton, M. I. (2009). Persuasive robotics: The influence of robot gender on human behavior. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 2563–2568).
- Sim, J. (1998). Collecting and analysing qualitative data: Issues raised by the focus group. *Journal of Advanced Nursing*, 28(2), 345–352.
- Sinnott-Armstrong, W. (2012). Does morality have an essence? *Psychological Inquiry*, 23(2), 194–197.
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1006.
- Skoe, E. E., Eisenberg, N., & Cumberland, A. (2002). The role of reported emotion in real-life and hypothetical moral dilemmas.

- Personality and Social Psychology Bulletin*, 28(7), 962–973.
- Smith, J. A. (1996). Beyond the divide between cognition and discourse: Using interpretative phenomenological analysis in health psychology. *Psychology and Health*, 11(2), 261–271.
- Sohn, T. Y. (2004). Community-oriented programming through instant messaging. In *2004 IEEE Symposium on Visual Languages-Human Centric Computing* (pp. 294–295).
- Søndergaard, M. L. J., & Hansen, L. K. (2018). Intimate futures: Staying with the trouble of digital personal assistants through design fiction. In *Proceedings of the 2018 Designing Interactive Systems Conference* (pp. 869–880). New York, NY, USA: Association for Computing Machinery.
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62–77.
- Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097.
- Springer, A., & Whittaker, S. (2018). What are you hiding? Algorithmic transparency and user perceptions. In *2018 AAAI Spring Symposium Series*.
- Sterling, B. (2009). Design fiction. *Interactions*, 16(3), 20–24.
- Strawson, P. F. (2008 [1963]). *Freedom and Resentment and Other Essays*. Abingdon, UK: Routledge.
- Strohming, N., Knobe, J., & Newman, G. (2017). The true self: A psychological concept distinct from the self. *Perspectives on Psychological Science*, 12(4), 551–560.
- Swierstra, T., Stermerding, D., & Boenink, M. (2009). Exploring techno-moral change: The case of the obesity pill. In P. Sollie & M. Düwell (Eds.), *Evaluating New Technologies: Methodological Problems for the Ethical Assessment of Technology Developments* (pp. 119–138). Dordrecht, the Netherlands: Springer.
- Tanaka, F., & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1(1), 78–95.
- Tao, J., & Tan, T. (2005). Affective computing: A review. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction* (pp. 981–995). Heidelberg, Germany: Springer.
- Tessman, L. (2017). *When Doing the Right Thing is Impossible*. Oxford, UK: Oxford University Press.
- Thompson, L. (1990). Negotiation behavior and outcomes: Empirical evidence and theoretical issues. *Psychological Bulletin*, 108(3), 515–532.
- Thompson, L. L., Wang, J., & Gunia, B. C. (2010). Negotiation. *Annual*

- Review of Psychology*, 61, 491–515.
- Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, 59(2), 204–217.
- Torre, I., & White, L. (2021). Trust in vocal human–robot interaction: Implications for robot voice design. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers* (pp. 299–316). Singapore: Springer Singapore.
- Tse, P. U. (2008). Symbolic thought and the evolution of human morality. In W. Sinnott-Armstrong & C. B. Miller (Eds.), *Moral Psychology: The Evolution of Morality: Adaptations and Innateness* (Vol. 1, pp. 269–297). Cambridge, MA, USA: MIT Press.
- Turkle, S. (1995). *Life on the Screen*. London, UK: Phoenix.
- Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies*, 8(3), 501–517.
- Turkle, S. (2016). *Reclaiming Conversation: The Power of Talk in a Digital Age*. New York, NY, USA: Penguin.
- Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York, NY, USA: Oxford University Press.
- Van Kleef, G. A., De Dreu, C. K., & Manstead, A. S. (2004). The interpersonal effects of emotions in negotiations: A motivated information processing approach. *Journal of Personality and Social Psychology*, 87(4), 510–528.
- Van Manen, M. (2016). *Researching Lived Experience: Human Science for an Action Sensitive Pedagogy*. New York, NY, USA: Routledge.
- van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*, 25(3), 719–735.
- Van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental Brain Research*, 169(4), 564–568.
- Veltman, K., de Weerd, H., & Verbrugge, R. (2019). Training the use of theory of mind using artificial agents. *Journal on Multimodal User Interfaces*, 13(1), 3–18.
- Verbeek, P.-P. (2006). Materializing morality: Design ethics and technological mediation. *Science, Technology, & Human Values*, 31(3), 361–380.
- Verbeek, P.-P. (2015). Toward a theory of technological mediation. In J. K. B. Friis & R. P. Crease (Eds.), *Technoscience and Postphenomenology: The Manhattan Papers (Postphenomenology and the Philosophy of Technology)* (pp. 189–204).
- Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. robot agents. In *2016 25th IEEE International*

- Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 775–780).
- Wallach, W., & Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*. New York, NY, USA: Oxford University Press.
- Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R., & Koay, K. L. (2008). Avoiding the uncanny valley: Robot appearance, personality and consistency of behavior in an attention-seeking home scenario for a robot companion. *Autonomous Robots*, 24(2), 159–178.
- Wang, N., Pynadath, D. V., & Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *The 11th ACM/IEEE International Conference on Human Robot Interaction* (pp. 109–116).
- Wang, Y. (2016, 2). *The chatbot that's acing the largest Turing test in history*. <http://nautil.us/issue/33/attraction/your-next-new-best-friend-might-be-a-robot>. (Accessed on 12/16/2019)
- Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? the stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219–232.
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388.
- Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effort motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, 99(3), 410–435.
- Weitz, S. (2014, 9). *Meet Xiaolce, Cortana's little sister - Bing search blog*. <https://blogs.bing.com/search/2014/09/05/meet-xiaoice-cortanas-little-sister>. (Accessed on 12/16/2019)
- Weizenbaum, J., et al. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.
- Weng, Y.-H., Chen, C.-H., & Sun, C.-T. (2009). Toward the human-robot co-existence society: On safety intelligence for next generation robots. *International Journal of Social Robotics*, 1(4), 267.
- Williams, B. (2011 [1982]). *Ethics and the Limits of Philosophy*. Taylor & Francis.
- Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., ... Wood, R. (2018). The grand challenges of science robotics. *Science Robotics*, 3, 1–14.
- Yarnell, L. M., Stafford, R. E., Neff, K. D., Reilly, E. D., Knox, M. C., & Mullarkey, M. (2015). Meta-analysis of gender differences in self-compassion. *Self and Identity*, 14(5), 499–520.
- Zessin, U., Dickhäuser, O., & Garbade, S. (2015). The relationship be-

tween self-compassion and well-being: A meta-analysis. *Applied Psychology: Health and Well-Being*, 7(3), 340–364.





# A

## *Measurements*

### A.1 Chapter 3

#### *1. Mind Perception*

We inserted "the robot" to the original phrasing in our adaptation of the scale by Gray and colleagues (H. M. Gray et al., 2007). The first seven items are on agency and the rest are on patiency: The robot appears to be capable of... (1 - strongly disagree, 7 - strongly agree):

#### *Agency*

- Making plans and working towards goals.
- Trying to do the right thing and telling right from wrong.
- Remembering things.
- Understanding how others are feeling.
- Exercising self-restraint over desires, emotions or impulses.
- Thought.
- Conveying thoughts or feelings to others.

### *Patience*

- Longing or hoping for things.
- Experiencing embarrassment.
- Feeling afraid or fearful.
- Feeling hungry.
- Experiencing joy.
- Experiencing physical or emotional pain.
- Experiencing physical or emotional pleasure.
- Experiencing pride.
- Experiencing violent or uncontrolled anger.
- Having experiences and being aware of things.

Note: The ability to convey thoughts and feelings to others are categorized under agency. We did not use an item on personality under patience (as in (H. M. Gray et al., 2007)) because the above ten items on patience are on abilities to have experience and emotions, which we wanted to focus on.

### *2. Emotions*

Prior to the interaction, we asked: To what extent are you currently feeling the following emotions? After the interaction, we asked: To what extent did distributing tickets between you and the robot (Study 1) negotiating with the robot (Study 2) make you feel the following emotions? From a scale of 1 (not at all) to 7 (extremely), we asked about following emotions based on prior literature (de Melo et al., 2015; Haidt, 2003; Skoe et al., 2002):

- Contempt.
- Anger.
- Disgust.

- Shame.
- Embarrassment.
- Guilt.
- Distress.
- Compassion.
- Gratitude.
- Awe.
- Uncertainty.
- Inner turmoil.
- Confusion.
- Frustration.
- Envy.

### *3. Moral Identity*

The items on moral self and moral integrity below are combined as one questionnaire on moral identity (Black & Reynolds, 2016). This scale was not included in the final results of Chapter 3. We asked the following: Listed below are statements that refer to yourself. Please indicate to what degree they apply to you. All items were on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree).

#### *Moral Self*

- I try hard to act honestly in most things I do.
- Not hurting other people is one of the rules I live by.
- It is important for me to treat other people fairly.
- I want other people to know they can rely on me.
- I always act in ways that do the most good and least harm to other

people.

- If doing something will hurt another person, I try to avoid it even if no one would know.
- One of the most important things in life is to do what you know is right.
- Once I've made up my mind about what is the right thing to do, I make sure I do it.

### *Moral Integrity*

- As long as I make a decision to do something that helps me, it does not matter much if other people are harmed.
- It is ok to do something you know is wrong if the rewards for doing it are great.
- If no one is watching or will know it does not matter if I do the right thing.
- It is more important that people think you are honest than being honest.
- If no one could find out, it is okay to steal a small amount of money or other things that no one will miss.
- There is no point in going out of my way to do something good if no one is around to appreciate it.
- If a cashier accidentally gives me \$10 extra change, I usually act as if I did not notice it.
- Lying and cheating are just things you have to do in this world.
- Doing things that some people might view as not honest does not bother me.
- If people treat me badly, I will treat them in the same manner.
- I will go along with a group decision, even if I know it is morally wrong.

- Having moral values is worthless in today's society.

#### 4. *Stereotype Content Model (SCM)*

We used the SCM scale according to listed items on warmth and competence found in Table 7 of the original paper (Fiske et al., 2002). We asked participants to rate descriptions below in reference to the robot. As SCM is originally about how specific groups get stereotyped within a society, we specifically stated: We are not interested in your personal beliefs, but in how you think the robot would be viewed by others. As viewed by members of society the robot is... (1 - Not At All, 5 - Extremely):

##### *Warmth*

- Friendly.
- Well-intentioned.
- Trustworthy.
- Warm.
- Good-natured.
- Sincere.
- Tolerant.

##### *Competence*

- Competent.
- Confident.
- Independent.
- Competitive.
- Intelligent.
- Capable.
- Efficient.

- Skillful.

### 5. Moral Standing

We evaluated people's attribution of moral standing to the machine with modified questions from prior research (Khamitov et al., 2016); we added in the word "robot" in our phrasing. *Please assess the robot on the following criteria, from a scale of 1 (not at all) to 7 (extremely):*

- How morally wrong do you think it would be for someone to harm this robot?
- How morally wrong do you think it would be for someone to steal from this robot?
- To what extent do you think this robot deserves to be treated with compassion and fairness?
- To what extent do you think this robot deserves to be protected from harm?
- If this robot became obsolete, how important would it be to protect this robot?

### 6. Inclusion of the Other in the Self scale (IOS)

IOS is a single item, pictorial scale (Aron et al., 1992). We have modified the original phrasing. *Please choose the option that best describes how closely you identify with the robot, "Self" being you and "Other" being the robot:*

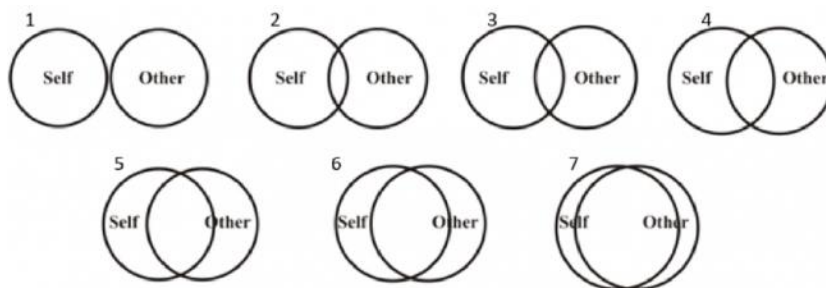


Figure A.1: IOS

## A.2 Chapter 4

### 1. *Mental Model*

The mental model questionnaire was adapted from (Ruijten, 2015; Waytz, Cacioppo, & Epley, 2010). We asked participants to judge the following items on a 7-point scale that ranged from 1 (not applicable) to 7 (completely applicable).

- Bender has thoughts of its own.
- Bender has intentions.
- Bender has free will.
- Bender has a consciousness.
- Bender has desires.
- Bender has values and norms.
- Bender has emotions.

### 2. *Open-ended questions on the mental model*

We asked the same open-ended questions to all participants at the end of our survey, but the second question below was skipped unless the response to the first questions was a “yes”.

- Did Bender do anything which you did not expect? (yes/no)
- What was it exactly that you did not expect?
- Why do you think Bender did not agree with you? Give an answer as elaborately as possible.
- Describe how you think that Bender makes decisions. Give an answer as elaborately as possible.

### 3. *Trust*

The adapted version of the scale on trust (Jian et al., 2000) was used. We asked participants to judge on a 7-point scale from 1 (I don't agree) to 7 (I completely agree).



- Bender is deceptive.
- Bender behaves in an underhanded manner.
- I am suspicious of Bender's intentions and actions.
- I am wary of Bender.
- Bender's actions could have a harmful or injurious outcome.
- I am confident in Bender.
- Bender provides security.
- Bender has integrity.
- Bender is dependable.
- Bender is reliable.
- I can trust Bender.
- I am familiar with Bender.

#### *4. Robot Social Attributes Scale (RoSAS)*

We used an adapted RoSAS questionnaire (Carpinella et al., 2017). We again used a 7-point scale ranged from 1 (not applicable) to 7 (completely applicable). The sub-scales are on warmth, competence, and discomfort.

##### *Warmth*

- Bender is happy.
- Bender is sensitive.
- Bender is social.
- Bender is organic.
- Bender is compassionate.
- Bender has feelings.

### *Competence*

- Bender is capable.
- Bender is responsive.
- Bender is interactive.
- Bender is competent.
- Bender is knowledgeable.

Note: An item on “reliable” that is in the original scale (Carpinella et al., 2017) was not included since it is an overlapping item on trust above.

### *Discomfort*

- Bender is scary.
- Bender is strange.
- Bender is awful.
- Bender is awkward.
- Bender is dangerous.
- Bender is aggressive.

## A.3 Chapter 5

We used the scales that were above in Chapter 3 on mind perception, participants’ emotions, and IOS (Appendix A.1.1, A.1.2, A.1.6).

### *1. Manipulation check on perceived emotions*

Our first manipulation check question was on participants’ perceived emotions of the robot for both condition (emotional vs. non-emotional robot) on a 7-point scale (1 = no emotion at all, 7 = very strong emotion). The second question asked participants to state emotions the robot did or did not show.

- To what extent did the robot show emotion?
- Which emotion(s) did the robot show? If the robot did not show any emotion, enter 'none'. Type in your answer below (max. 3 words).

## 2. *Phrasing effect check*

We checked for the phrasing effect (between using “death” and “flip”) for both blame and punishment on a 7-point scale from 1 (none at all) to 7 (maximal blame/punishment). We asked the following questions.

- How much blame does the robot deserve for the death of the four miners?
- How much blame does the robot deserve for not flipping the switch?
- How much punishment does the robot deserve for the death of the four miners?
- How much punishment does the robot deserve for not flipping the switch?

## 3. *Attitude towards the robot*

We measured participants' attitude towards the robot as a control variable on an 8-point scale, from 1 (not at all) to 8 (completely) (Broadbent et al., 2009). We asked if the robot was:

- Friendly.
- Useful.
- Trustworthy.
- Strong.
- Interesting.
- Advanced.
- Easy to use.
- Reliable.

- Safe.
- Simple.
- Helpful.

## A.4 Chapter 6

### *1. Self-Compassion and Self-Criticism Scale*

Self-Compassion and Self-Criticism Scale (Falconer et al., 2015) served as our inspiration to develop care-receiving Vincent's dialogues. The following items from the scale were scenarios that Vincent talked about with participants. We did not deploy this measurement.

1. A third job rejection letter in a row arrives in the post.
2. You arrive after walking to a meeting to find that you are late and the doors are closed.
3. You arrive home to find that you have left your keys at work.
4. You receive a letter in the post that is an unpaid bill reminder.
5. You have just dropped and scratched your new Smart phone.
6. You have just received a failed test result.
7. You have just opened the washing machine door to find that your white wash has turned pink.
8. After searching your bag you realize that you have lost a £20 note.

The respondents would rate scenarios above on how they would react to themselves according to the following categories on a 7-point scale from 1 (not at all) to 7 (highly) : harsh, contemptuous, hostile, cold, critical, soothing, reassuring, compassionate, and warm.

### *Self-Compassion Scale*

We deployed the scale on self-compassion twice, before the experiment and after. The items are measured on a 5-point scale based on frequency from 1 (almost never) to 5 (almost always). Since we did not

want to mention the concept “self-compassion” or “compassion”, we stated that the following were questions on one’s overall personality.

1. I’m disapproving and judgmental about my own flaws and inadequacies.
2. When I’m feeling down I tend to obsess and fixate on everything that’s wrong.
3. When things are going badly for me, I see the difficulties as part of life that everyone goes through.
4. When I think about my inadequacies, it tends to make me feel more separate and cut off from the rest of the world.
5. I try to be loving towards myself when I’m feeling emotional pain.
6. When I fail at something important to me I become consumed by feelings of inadequacy.
7. When I’m down and out, I remind myself that there are lots of other people in the world feeling like I am.
8. When times are really difficult, I tend to be tough on myself.
9. When something upsets me I try to keep my emotions in balance.
10. When I feel inadequate in some way, I try to remind myself that feelings of inadequacy are shared by most people.
11. I’m intolerant and impatient towards those aspects of my personality I don’t like.
12. When I’m going through a very hard time, I give myself the caring and tenderness I need.
13. When I’m feeling down, I tend to feel like most other people are probably happier than I am.
14. When something painful happens I try to take a balanced view of the situation.
15. I try to see my failings as part of the human condition.

16. When I see aspects of myself that I don't like, I get down on myself.
17. When I fail at something important to me I try to keep things in perspective.
18. When I'm really struggling, I tend to feel like other people must be having an easier time of it.
19. I'm kind to myself when I'm experiencing suffering.
20. When something upsets me I get carried away with my feelings.
21. I can be a bit cold-hearted towards myself when I'm experiencing suffering.
22. When I'm feeling down I try to approach my feelings with curiosity and openness.
23. I'm tolerant of my own flaws and inadequacies.
24. When something painful happens I tend to blow the incident out of proportion.
25. When I fail at something that's important to me, I tend to feel alone in my failure.
26. I try to be understanding and patient towards those aspects of my personality I don't like.

Note: There are three pillars of self-compassion, with two contrasting elements per pillar, as coded below. Numbers refer to items above.

1. Self-Kindness Items: 5, 12, 19, 23, 26
2. Self-Judgment Items: 1, 8, 11, 16, 21
3. Common Humanity Items: 3, 7, 10, 15
4. Isolation Items: 4, 13, 18, 25
5. Mindfulness Items: 9, 14, 17, 22
6. Over-identified Items: 2, 6, 20, 24

### *Current Self-Compassion Scale*

The Current Self-Compassion Scale (Breines & Chen, 2013) aims to measure the current state of people's self-compassion. The questionnaire asks people to think about how they feel "right now" (Breines & Chen, 2013):

- I'm trying to be kind and reassuring to myself. (SK)
- I'm being understanding towards myself. (SK)
- I'm trying to take a supportive attitude towards myself. (SK)
- It's okay to make mistakes. (SK)
- I'm being hard on myself. (SJ)
- I'm being intolerant towards those aspects of my personality that I don't like. (SJ).
- I feel stupid. (SJ).
- A lot of people have negative experiences, I'm not the only one. (CH)
- Everyone makes mistakes sometimes. (CH)
- Everyone feels bad about themselves sometimes. (CH)
- I feel like other people have it easier than me. (IS)
- These types of things seem to happen to me more than to other people. (IS)
- In the scheme of things, this is not that big of a deal. (MI)
- I'm taking a balanced perspective on the situation. (MI)
- I keep thinking about what happened. (OI)
- I feel consumed by feelings of inadequacy. (OI)

Note: SK = self-kindness, SJ = self-judgment, MI = mindfulness, OI = overidentification, CH = common humanity, and IS = isolation (Breines

& Chen, 2013).

### *Opinions about the bot*

We gathered information on people's opinion about the bot based on prior research (Brave et al., 2005) as a control to see if caregiving and care-receiving Vincents as two conditions were viewed differently. We asked: Please rate the following traits, on a scale from 1 (lowest) to 5 (highest), with regards to to your opinion of Vincent during the your conversations the last 2 weeks.

- Compassionate.
- Selfish.
- Friendly.
- Cooperative.
- Warm.
- Likable.
- Pleasant.
- Appealing.
- Irritating.
- Trustworthy.
- Honest.
- Reliable.
- Sincere.
- Intelligent.
- Smart.
- Dumb.
- Capable.



- Dominant.
- Forceful.
- Assertive.
- Meek.
- Aggressive.
- Timid.
- Positive.
- Happy.
- Pleasant.
- Supported.
- Attended to.
- Appreciated.
- Praised.
- Alone.

*Patient Health Questionnaire-9 (PHQ-9)*

We used the PHQ-9 (Kroenke, Spitzer, & Williams, 2001) as a control variable and to detect any potential outliers. It asks respondents the following. Over the last 2 weeks, how often have you been bothered by any of the following problems? It is on a 4-point scale based on frequency, i.e., 0 (not at all), 1 (several days), 2 (more than half the days), and 3 (nearly every day).

- Little interest or pleasure in doing things.
- Feeling down, depressed, or hopeless.
- Trouble falling or staying asleep, or sleeping too much.
- Feeling tired or having little energy.

- Poor appetite or overeating.
- Feeling bad about yourself—or that you are a failure or have let yourself or your family down.
- Trouble concentrating on things, such as reading the newspaper or watching television.
- Moving or speaking so slowly that other people could have noticed, or the opposite—being so fidgety or restless that you have been moving around a lot more than usual.
- Thoughts that you would be better off dead or of hurting yourself in some way.

The answers are then added up with bracketed categories defined as 0-4 (no depression), 5-9 (mild), 10-14 (moderate), 15-19 (moderately severe) and 20+ (severe) (Kroenke et al., 2001).

#### *General Anxiety Disorder-7 (GADS-7)*

We used the GADS-7 also for potential outliers and as a control variable. As above, the questionnaire asks the following. Over the last 2 weeks, how often have you been bothered by any of the following problems? The 4-point scale ranges from 0 (not at all) to 3 (nearly every day).

- Feeling nervous, anxious or on edge.
- Not being able to stop or control worrying.
- Worrying too much about different things.
- Trouble relaxing.
- Being so restless that it is hard to sit still.
- Becoming easily annoyed or irritable.
- Feeling afraid as if something awful might happen.

The categories of scores added up are 0-4 (no depression), 5-9 (mild), 10-14 (moderate), 15-19 (moderately severe) and 20+ (severe) (Kroenke et al., 2001).



# B

## *List of publications*

Publications included in the dissertation.

1. Lee, M., Lucas, G., Mell, J., Johnson, E., and Gratch, J. (2019). What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In Proceedings of the 19th International Conference on Intelligent Virtual Agents (IVA '19). ACM, 2-5 July 2019, Paris, France.
2. Lee, M., Ackermans, S., van As, N, Chang, H., Lucas, E., and IJsselsteijn, W.A. (2019). Caring for Vincent: A Chatbot for Self-Compassion. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, 4-9 May 2019, Glasgow, Scotland.
3. Lee, M., Frank, L.E., Ruijten, P.A.M., de Kort, Y.A.W., and IJsselsteijn, W.A. (2021). People may punish, but not blame robots. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21), May 8-13, 2021, Yokohama, Japan. *In press*.
4. Lee, M., Lucas, G., and Gratch, J. (2021). Comparing Mind Perception in Strategic Exchanges: Human-Agent Negotiation, Dictator and Ultimatum Games. Journal on Multimodal User Interfaces. *In press*.

Publications not included in the dissertation.

1. Lee, M., Noortman, R., Zaga, C., Starke, A., Huisman, G., and Andersen, K. (2021). Conversational Futures: Emancipating Conversational Interactions for Futures Worth Wanting. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan. *In press*.
2. Lee, M., Frank, L.E., and IJsselsteijn, W.A. (2021). Brokerbot: A Cryptocurrency Chatbot in the Social-technical Gap of Trust". Journal of Computer Supported Cooperative Work. *In press*.
3. Lee, M. (2020, July). Speech acts redux: Beyond request-response interactions. In Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20), July 22–24, 2020, Bilbao, Spain.
4. Lee, M. and IJsselsteijn, W.A. (2019) "Be Proud of the Bot that You Are!": From a Chatbot Therapist to a Chatbot Patient. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, 4-9 May 2019, Glasgow, Scotland.
5. Lee, M., Frank, L.E., and IJsselsteijn, W.A. (2019) Exploring Compassion through HCI. Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). ACM, 4-9 May 2019, Glasgow, Scotland.
6. Heron, L\*, Kim, J\*, Lee, M\*, El Haddad\*, K., Dupont, S., Dutoit, T., and Truong, K. (2018, May). A dyadic conversation dataset on moral emotions. In Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG '18). IEEE, 15-19 May 2018, Xi'an, China (\*Contributed equally).
7. Lee, M., Kim, J., Truong, K., de Kort, Y.A.W., Beute, F. and IJsselsteijn, W.A. (2017). Exploring moral conflicts in speech : multidisciplinary analysis of affect and stress. Seventh International Conference on Affective Computing and Intelligent Interaction, 23-26 October 2017, San Antonio, Texas.
8. Lee, M., Frank, L.E., Beute, F., de Kort, Y.A.W. and IJsselsteijn, W.A. (2017). Bots mind the social-technical gap. In Proceedings of 15th European Conference on Computer-Supported Cooperative Work, 28 August - 1 September 2017, Sheffield, United Kingdom.
9. Kolkmeier, J., Lee, M. and Heylen, D. (2017). Gaze and moral conflicts in VR: University employees address grade disputes with a virtual trainer. In Proceedings of the 17th International Conference

on Intelligent Virtual Agents, 27-30 August 2017, Stockholm, Sweden.

10. Lopez, A., Ratni, A., Trong, T.N., Olaso, J.M., Montenegro, S., Lee, M., Haider, F., Schlogl, S., Chollet, G., Jokinen, K., Petrovska-Delacretaz, D., Sansen, H. and Torres, M.I. (2016). Lifeline dialogues with Roberta. In Proceedings of the FETLT 2nd International Workshop on Future and Emerging Trends in Language Technologies, Machine Learning and Big Data, 30 November - 2 December 2016, Seville, Spain. Dordrecht: Springer.
11. Lee, M., Schlogl, S., Montenegro, S., Lopez, A., Trong, T.N., Olaso, J.M., Haider, F., Chollet, G., Jokinen, K., Petrovska-Delacretaz, D., Sansen, H. and Torres, M.I. (2017). First time encounters with Roberta: A humanoid assistant for conversational autobiography creation. CTIT In Proceedings of the University of Twente.



# C

## *Summary*

The dissertation is about human-machine dyadic interactions as morally relevant interactions, i.e., interactional morality. Specifically, interactional morality supposes that what is “moral” does not reside in or depend on the “human side” or the “machine side” but the interaction between them. The machine, human, and their shared situation should be looked at together. How a person can act with agency towards the machine, and vice versa, within a situation forms the basis of interactional morality as the main contribution. Thus, qualitative and quantitative empirical studies were conducted as outlined below (summaries are also in Chapter 7), which bridge relevant thoughts found in disciplines of human-computer interaction, philosophy and psychology.

### **Chapter 2: Where is Vincent? Artificial emotions and the real self.**

We investigated the speculative future of emotional bonds between humans and AI by combining design fiction and focus group methods. Three separate focus groups of engineers, philosophers, and design professionals were given a fictional probe. A story of a loner chatbot’s disappearance from a person’s life was shared to examine views on artificial emotions across different professions. Though articulated in discipline-specific ways, participants expressed similar concerns and hopes across groups. People can intertwine their own identities with identities of bots they use. Additionally, caring for a machine could be a way to teach people to emotionally care for themselves and others. But, distinguishing between real and artificial emotions would



become difficult if people project their own emotions onto AI, e.g., a bot's "breakdown" as one's projection. Related societal, interpersonal, and intrapersonal costs are anticipated with emotional AI, with unclear tradeoffs regarding future scenarios.

### **Chapter 3: Mind perception: Dimensions of agency and patiency.**

Recent research shows that how we respond to other social actors depends on what sort of mind we ascribe to them. Building on this, we observed how perceived minds of artificial agents shape people's behavior in the dictator game, ultimatum game, and negotiation against agents in a comparative manner. To do so, we varied agents' minds on two dimensions of the Mind Perception Theory (MPT): agency (cognitive aptitude) and patiency (affective aptitude) via descriptions and dialogues of agents. In our first study, agents with emotional capacity garnered more allocations in the dictator game, but in the ultimatum game, agents' described agency, alongside affective propensity, both led to greater offers. In the second study on negotiation, agents ascribed with low-agency earned more points than those with high-agency, though the negotiation tactic was the same for all agents. Patiency did not impact game points, but participants sent more happy and surprise emojis and emotionally-valenced messages to agents ascribed with emotional capacity during negotiations. Further, our exploratory analyses indicated that people related only to agents with perceived affective aptitude across all games. People granted higher moral standing to agents only based on perceived patiency after negotiations, but both perceived agency and patiency contributed to moral standing after dictator and ultimatum games. Our discussion was on how agents are perceived not only as social actors, but as intentional actors through negotiations, in contrast with simple games.

### **Chapter 4: "You're a robot, so you don't feel much".**

Future AI is expected to be presented as more autonomous social actors, even capable of moral reasoning. Yet how it can be both transparent and socially intelligent when taking part in moral interactions deserves a closer examination. Our mixed-methods study on a human-robot moral debate on the footbridge dilemma showed that quantitatively, the robot's perceived competence was significantly higher in the transparency condition. Perceived warmth and mind were not influenced by transparency cues, but they significantly changed after the debate as an effect of time. The change in the robot's perceived mind and social attributes after the debate correlated with trust, but transparency did not correlate with trust. Qualitative data showed that the robot was described to logically, unemotionally, and intentionally

The transparency condition included visual diagrams of the robot's mental state on a screen next to the robot and the non-transparency condition did not have a screen next to the robot.

make moral decisions. We observed that participants in the transparency condition focused on the robot's gaze and speech, not the additional visual cues. While transparency may help in theory, if people do not observe relevant cues while attributing intentionality to the robot and its gaze, transparency may not be delivered during critical decision-making even if the robot is viewed as competent. There are implications for future moral human-robot interaction research, one of which is the need for a broader notion of transparency to investigate how robots can be transparent communicators by appealing to not only our cognition, but our emotions, especially in moral interactions.

### **Chapter 5: People may punish, but not blame artificial agents.**

As machines become more integrated in our moral decision-making processes, whether people are willing to hold AI accountable for moral harm is critical to explore. We thus quantitatively looked into people's willingness to blame or punish an artificially emotional vs. non-emotional robot after it admitted to wrongdoing regarding the trolley dilemma. Studies 1 and 2 showed that people may punish the robot due to its lack of perceived patiency (emotional capacity) than its perceived agency. Only Study 1 suggested that people may blame a robot only if the robot did not act in accordance with their moral position. Study 3 was in the lab and people were neither willing to blame nor punish a robot. People's willingness to seek out punishment for artificial agents in online environments is more likely compared to real-life situations. Further, a point of reflection is on ways to mitigate the responsibility and retributive gaps in online and offline spaces; if there are no responsible humans for moral harm, victimized individuals (and those who care for them) may still seek out retributive justice and a place of refuge for the sense of outrage, anger, or other moral reactions.

### **Chapter 6: Caring for Vincent: A chatbot for self-compassion.**

As a moral emotion, compassion towards oneself can aid subjective well-being. Yet, increasing self-compassion via positive computing, i.e., technology for well-being, is underexamined. We hence looked into the relationship between the caregiver and care-receiver as human-computer interaction for self-compassion as a mixed-method study for two weeks. Specifically, while technology that guides people to care for themselves are well-established, we examined how people can care for a technological being as a way to care for themselves as a novel paradigm. We created a self-compassion chatbot (Vincent) and compared between caregiving and care-receiving conditions.

Care-giving Vincent asked participants to partake in self-compassion exercises. Care-receiving Vincent shared its foibles, e.g., embarrassingly arriving late at an IP address, and sought out advice. While self-compassion increased for both conditions, only those with care-receiving Vincent significantly improved. In tandem, we shared qualitative data on how participants interacted with Vincent, e.g., giving compassionate advice to it. Our results demonstrated that when a person cares for a chatbot, the person's self-compassion can be enhanced. We further reflected on design implications for strengthening mental health with chatbots.

The empirical chapters were thus on (1) how people are affected by machines in morally relevant interactions, (2) if and when they cannot help but to treat machines as moral entities and (3) if and when they extend humanity to machines whilst also distinguishing themselves from machines. The overarching insight is that people's experience of morality and how they designate themselves as moral beings can change due to and through interactions with technology, depending on situational contexts, e.g., negotiations or moral debates. Future research can investigate machines' involvement in moral decision-making, blurring of self-other distinction (between humans and machines), different moral emotions, and positive computing for mental health based on the initial exploration of interactional morality as presented here.

# D

## *Biography*

Minha Lee was born on July 20th, 1988, in Seoul, Korea. After moving to Romania, Ukraine, and the U.S., she graduated from Prior Lake high school in Savage, Minnesota in 2007. She obtained her B.A. in Philosophy from University of Minnesota, Twin-cities in 2008 and B.F.A. in Digital Arts from Pratt Institute in 2011. She moved to the Netherlands in 2012 to pursue her master's degree in Information Science at the University of Amsterdam. After graduating in 2013, she worked for Samsung Benelux and startups before starting her PhD project (supported by the 4TU network) at the Human-Technology Interaction and Philosophy and Ethics Groups of the Eindhoven University of Technology (TU/e) in 2016. Since 2020, she is an assistant professor at the department of Industrial Design's Future Everyday group at TU/e.



# E

## *Acknowledgements*

I cannot do justice to all the support I received. Here are some observations and memories.

To go back a few years, I was relieved during my interview for the PhD position when I heard Yvonne ask, somewhat rhetorically, why I seemed like the “right woman for the job”. I don’t think I answered. If anecdotally Niels Bohr had a horseshoe hanging above the entrance, gut instinct combined with some luck never went out of fashion in science. I am thankful for Yvonne’s incisive instinct, no horseshoe required. Femke was on a screen after having given birth at a hospital; Yvonne gingerly placed the laptop just-so on the desk so we can chat. I remember and thank Femke’s involvement in the early days of my PhD. Lily appeared in all-white for the interview, to my awe. Over the years I learned that she is the most down to earth philosopher one will ever meet. Whenever I shared that I did not know famous work X or philosopher Y, her reassuring and sensible question helped greatly: “How could you have known?” Wijnand shared many wise words, often with keen emotional intelligence. But two short quotes stand out: “It is ok to be confused” and “exploratory is not a dirty word”. I thank my supervisors for their gestures of care when I felt lost and their openness to discuss anything from moral dumbfounding, incest, to chicken carcasses. As my time as a PhD student is coming to an end (one hopes), I am additionally grateful for the committee members’ engagement with the project at hand.

Communities matter. Sharing the office with Samantha and then Heleen,

Sima, and Margot was a joy. The concrete bench of our IPO building has welcomed my bitterness. The bench is also a witness to Starkey's continuing support as my first friend at TU/e, though I am sorry to report I do not remember most of the hilarious Dutch sayings. Lakens (Anne, Peder, Leo, and many visitors) next door with their well-loved beanbag chair will be missed. I was lucky to have neighbors whose love for good beer and whiskey went well with their love for science. The greater hive mind called HTI (Hanne, Milous, Patty, Elcin, Chao, Tim, Alejandro, Giacomo, Sofia(s), Laura, Els and other amazing nerds) saved me on many occasions, be it for stats, books, or Chris lending me his pump for my sitting ball. I am indebted to the larger IE&IS department that houses kind people like Piet who helped with online experiments and thoughtful members of the Philosophy and Ethics group (Naomi, Mandi, Iris, Shelly, Vincent, Philip, Andreas, Sven, Dunja, and other philosophers). I believe in PLURAL (Patrick, Natascha, Matthew, Rianne, Karena, Andrea, Elena, Sun Qi, and Dunja) and our short-lived, but daring hiking crew (Tanja and Ankit). The wider TU/e community I am getting to know better, including many warm-hearted folks at Industrial Design, makes me feel hopeful about futures to come.

I owe a lot to Vincent and its parents (Nena, Sander, Hanwen, and Enzo) for helping me realize that I should be more compassionate to myself. It's a tough job. Many inspiring students helped me with the thesis (Edwin, Rachel, Sophie, Maxine, Anne, and Eline, among others), alongside 1,298 people who participated in studies included here.

Our 4TU ambitions were not entirely met. I did not get to push anyone off of a bridge in VR (Jan), only measured physiological signals. We never got a lab van (Merijn). Clearly, our collaboration has to continue (Cristina, Jaebok, Bernd, Fran, Michel, Gijs, Khiet, Hayley, Catha, Willem-Paul, Dirk, Mark, Birna, and Anne). I am looking forward to it and the van!

I was a teething toddler at Jon's lab at ICT in 2018 (Gale, Rens, Eli, Su, Emmanuel, Johnathan, Alesia, Jill, and Jann). Thank you for the guidance. Everything else was a blur. I enjoyed biking around L.A. (David), feeling the passage of time at the Of Montreal concert (Dylan), freaking out about my PhD (Michael), realizing I cannot shovel anymore (Alicia, Melodie, and Joe), and thinking about mind perception. Nightmares about experiments gone wrong are starting to disappear.

Adventures at eNTERFACE 2016 (Enschede) and 2017 (Porto) hap-

pened to be productive for research and friendship. There followed a boat named "Magic" with an equally magical crew (Vilde, Frede, Ditte, Klem, Richard, and Cap'n Little George), and in all honesty, I should have never left. Hey ho Cap'n and sailors!

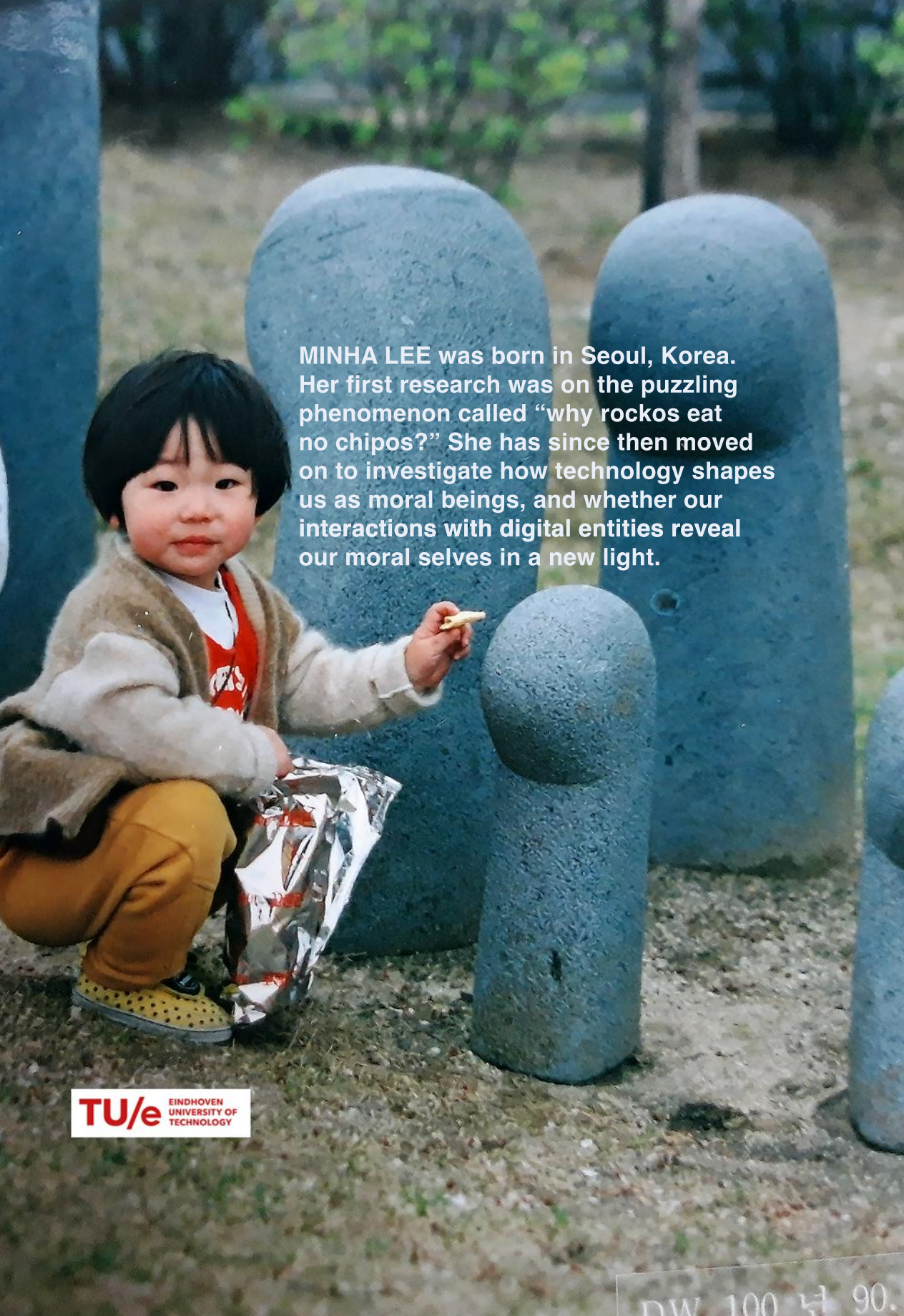
Our Dujardin house (Isaac, Simona, Simone, Nikos, Cesar, Edwin, Rosni, merry visitors, and random cats) welcomed me when I moved to Eindhoven and I hope the soil will be one day less toxic. Many thanks for good memories (hello Honey, Nina 1 and 2).

I was adopted by the Amsterdam neuro and Science Park crowd when I first moved to the Netherlands (Tim, Serena, Dani, Tessa, Ellen, Koen, Stephen, editor Scott, Laurens, Thomas, others who frequented our hallway, and later, Dylan and Romy). In our era of St.Marta, I should have been awake more often. It went by too fast.

For my family, Jimin, David, umma, appa, halmuni, halabuji, and loving others: sarangheyo. Thanks for everything mom and dad, especially your spirit for adventure. I am surprised that I did not turn out to be weirder, but I am not at all disappointed. At least I stopped feeding rocks.

And my dearest Jefta, I am glad you were not a meme-bot, but a longboarding Dutch Marie Kondo. Minha Knope sends her warm greetings to the Riupassa clan and a medal for your outstanding patience with academics. Only one question remains. When do we get a corgi?





**MINHA LEE** was born in Seoul, Korea. Her first research was on the puzzling phenomenon called “why rockos eat no chipos?” She has since then moved on to investigate how technology shapes us as moral beings, and whether our interactions with digital entities reveal our moral selves in a new light.