

Turing Tests as Reflexive Experimental Apparatus

by

Joshua David Bruce Redstone

A thesis submitted to the Institute of Cognitive Science in partial
fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Cognitive Science

Carleton University
Ottawa, Ontario

©2019, Joshua David Bruce Redstone

Abstract

Over half a century ago, Alan Turing proposed “the Imitation Game” as a test of whether machines such as digital computers can be said to think. Subsequent discussion of Turing tests – human-machine interactions that are importantly similar to Turing’s original Imitation Game – has been limited in its understanding of what they are good for, viewing them as good either for prompting philosophical reflection on the limits of our concept of the mental or for addressing empirical questions about the machines involved in human-machine interactions. This project is an attempt to expand our understanding of what Turing tests are good for: my novel proposal is that they are good for addressing empirical questions about the humans involved in human-machine interactions. More simply put, I argue Turing tests are useful not merely as conceptual prompts or non-reflexive experimental apparatus, but as reflexive experimental apparatus. I begin with an examination of Turing’s own work and of the subsequent discussion’s limited understanding of the usefulness of Turing tests as either conceptual prompts or non-reflexive experimental apparatus. I then lay out the key elements of my novel proposal that they are useful as reflexive experimental apparatus. Finally, I offer a “proof of concept” for this novel proposal by describing and discussing the results of one preliminary attempt to use Turing tests as reflexive experimental apparatus.

Acknowledgements

I owe many thanks to Dr. David Matheson, who has patiently guided me while I worked out many of the conceptual and theoretical parts of this project, and whose guidance and advice have been indispensable throughout my graduate student career. Dave also helped me refine many of these ideas, helping me to envision a more precise research project. My thanks also to Dr. Chris Herdman, who took an interest in this project, and provided me the opportunity to bring it to life in the ACE Lab at Carleton University. In short, Chris gave me a place to work, and put me in touch with the right people to work with. One of those people, Dr. Kathy Van Bentham, graciously assisted me and my experimental team with the analysis of our data, the writing of our protocol forms, and many of the other “nuts and bolts” of the project. Her assistance is especially appreciated with these matters, because I’m a philosopher by training, not a scientist. But thanks to the guidance of these three individuals, I feel as though I’m becoming a decent scientist, and continuing to develop as a philosopher as well.

Speaking of my experimental team, I had the pleasure of working with a fantastic group of people. I’m very proud of what we achieved together, as I’m sure all of my team members are. James Howell and Kevin Hua brought our experimental apparatus to life. They created a wonderful interface with which we could study empathy and mind perception with our virtual characters. I am indebted to them for their work on this part of the project, which was absolutely essential to our team’s success. Kevin also helped the team to run the experiment, as did Clare Coyle, Melanie Turabian and Ahmad Hachem. Clare, Melanie and Ahmad also helped to design the experiment. Their ideas contributed

to making our experiment even more comprehensive than it would have been had I designed it alone. Each of them has my sincerest thanks.

I also wish to thank Cassandra Ommerli for her early contributions to the project. Cassandra did not work with us for long, but her early contributions to the team are much appreciated. I must also thank Dr. Jim Davies for agreeing to sit on my thesis committee, along with Dave and Chris. Of course, I cannot forget to thank my external examiner, Dr. Dan MacArthur, and my internal examiner, Dr. Michael Mopas. Jim's, Dan's and Michael's participation in my thesis examination made for an engaging and interesting discussion. Thanks to them, I have a whole new set of questions to explore now that this project is complete. My thanks also go out to the other members of the ACE lab, whose questions and feedback during our weekly lab group meetings helped my experimental team and I to further hone our ideas for this project. Likewise, I want to thank to the students and faculty of the Institute of Cognitive Science at Carleton University, who also provided very helpful feedback, and who asked some very illuminating questions about this project, at the 2018 ICS Spring Conference.

Finally, I want to say thank you to those, without whose past help, I would never have managed to make it this far. Dr. Wayne Borody of Nipissing University, who first got me seriously thinking about philosophical and scientific questions that surround humans and machines when I was an undergraduate – questions that would eventually culminate in research like this – has my eternal gratitude. Thanks also to Dr. Heidi Maibom, who helped me to develop my ideas in my early years at graduate school. Last, but certainly not least, I want to thank my family: my father Harold, my mother Yvonne and my sister Michelle kept me on track throughout all my years in graduate school with

all kinds of indispensable support. And thanks to my partner Sara, who is a light in the darkness, the north star that shines above rough seas, and my most steadfast friend.

Table of Contents

Abstract.....	1
Acknowledgements	2
Table of Contents	5
List of Tables	8
List of Figures.....	11
List of Appendices.....	12
Introduction.....	13
Chapter One: On What Turing Tests Are Good For	18
1.1 The Imitation Game.....	20
1.2 On Thinking Machines	24
1.3 Turing Tests as Conceptual Prompts	37
1.4 Turing Tests as Non-reflexive Experimental Apparatus	41
1.5 Toward Re-Imagining What Turing Tests are Good For	44
Chapter Two: On the Usefulness of Turing Tests as Reflexive Experimental Apparatus	48
2.1 Levels of Indistinguishability and Harnad’s Turing Test Hierarchy	50
2.2 Ishiguro’s Turing Tests.....	54
2.3 Empirical Questions about the Humans Involved in Human-Machine Interactions: Empathy, Mind Perception and Mindreading.....	56
2.4 Toward A “Proof-of-Concept” of Turing Tests as Reflexive Experimental Apparatus.....	65
Chapter Three: Using Turing Tests as Reflexive Experimental Apparatus	68
3.1 Research Questions	70
3.2 Methods	73
3.2.1 Participants.....	74

3.2.2	Materials.....	75
3.2.2.1	Personnel	75
3.2.2.2	Experimental Stimuli.....	78
3.2.2.3	Experimental Apparatus	81
3.2.2.4	Measures.....	84
3.2.2.4.1	Self-Report Measures	84
3.2.2.4.2	Physiological Measures	88
3.2.3	Procedure	89
3.2.3.1	Pre-Interaction	89
3.2.3.2	Interaction.....	91
3.2.3.3	Post-Interaction	95
3.3	Results	96
3.3.1	Sample Characteristics	96
3.3.2	Experimental Results	101
6.4	Discussion.....	119
	Conclusion	126
	Appendices.....	131
	Appendix A - Virtual Agent Scripts	131
A.1	Robotic Agent Script.....	131
A.2	Humanoid Agent Script.....	140
	Appendix B - Questionnaire Items	149
B.1	Pre-test Questionnaire Items	149
B.2	Empathy Questionnaire Items	150
B.3	NARS Questionnaire Items	151
B.4	Post-test Questionnaire Items.....	152
	Appendix C - Protocol Forms.....	155

C.1	Informed Consent Form	155
C.2	Debriefing Form.....	160
C.3	SONA Posting.....	162
Appendix D - “Twenty Questions” Item Descriptions		163
D.1	Inanimate Objects:	163
D.2	Animate Objects:.....	163
Appendix E - Experimenter’s Reference Materials.....		164
Appendix F - Ethical Clearance.....		166
References		168

List of Tables

Table 1: Experience with Virtual Reality and Robots	97
Table 2: Total Mean NARS Scores	97
Table 3: Total Mean BES-A Scores.....	98
Table 4: Group Means – Experience with the Virtual Agent	100
Table 5: Mean Perceived Experience Ratings	102
Table 6: Component Loadings for Perceived Experience	102
Table 7: Mean Perceived Agency Ratings.....	102
Table 8: Component Loadings for Perceived Agency	102
Table 9: Group Means for Perceived Experience and Perceived Agency	103
Table 10: Group Means – Turing Test Questions.....	105
Table 11: Cut-Points for Cognitive and Affective BES-A Scores.....	108
Table 12: Mean Perceived Experience Ratings of Low & High Cognitive Empathy Groups.....	108
Table 13: Tests of Between-Subject Effects – Dependent Variable: Perceived Experience	109
Table 14: Bayesian ANOVA – Model Comparison: Perceived Experience	109
Table 15: Mean Perceived Agency Ratings of Low & High Cognitive Empathy Groups	109
Table 16: Tests of Between-Subject Effects – Dependent Variable: Perceived Agency	110
Table 17: Bayesian ANOVA – Model Comparison: Perceived Agency	110
Table 18: Mean Perceived Experience Ratings of Low & High Affective Empathy Groups.....	111

Table 19: Tests of Between-Subject Effects – Dependent Variable: Perceived Experience	111
Table 20: Bayesian ANOVA – Model Comparison: Perceived Experience	111
Table 21: Mean Perceived Agency Ratings of Low & High Affective Empathy Groups	112
Table 22: Tests of Between Subject Effects – Dependent Variable: Perceived Agency	112
Table 23: Bayesian ANOVA – Model Comparison: Perceived Agency	112
Table 24: Low & High Affective Empathy Group Mean Agreement Ratings for “The character was probably controlled by a person.”	113
Table 25: Tests of Between Subjects Effects – Dependent Variable: “The character was probably controlled by a person.”	114
Table 26: Bayesian ANOVA – Model Comparison: “The character was probably controlled by a person.”	114
Table 27: Low & High Cognitive Empathy Group Mean Ratings for “The character was probably controlled by a person.”	114
Table 28: Tests of Between-Subjects Effects – Dependent Variable: “The character was probably controlled by a person.”	115
Table 29: Bayesian ANOVA – Model Comparison – “The character was probably controlled by a person.”	115
Table 30: Low & High Affective Empathy Group Mean Ratings for “The character was probably controlled by an artificial intelligence.”	116
Table 31: Tests of Between Subjects Effects – Dependent Variable: “The character was probably controlled by an artificial intelligence.”	116

Table 32: Bayesian ANOVA – Model Comparison – “The character was probably controlled by an artificial intelligence.”	116
Table 33: Low & High Cognitive Empathy Group Mean Agreement Ratings for “The character was probably controlled by an artificial intelligence.”	117
Table 34: Tests of Between-Subjects Effects – Dependent Variable: “The character was probably controlled by an artificial intelligence.”	117
Table 35: Bayesian ANOVA – Model Comparison: “The character was probably controlled by an artificial intelligence.”	118

List of Figures

Figure 1: The experimenter administers the experiment with the participant.	77
Figure 2: Members of the experimental team acting as the puppetmaster. One is using the 3D interface (left) and the other is using the 2D interface (right).	77
Figure 3: The AltspaceVR avatar chosen for V-2's physical appearance (above) and Veronica's physical appearance (right).	79
Figure 4: Oculus Rift CV-1 Headset.....	81
Figure 5: The 3D interface (left) and the 2D interface (right) from the puppetmaster's perspective	82
Figure 6: The Empatica E4 Wristband.....	88
Figure 7: Distribution of Total NARS Scores.....	98
Figure 8: Distribution of Total BES-A Scores.....	99
Figure 9: Participants' Turing Test Judgements	105
Figure 10: Low and High Cognitive Empathy Group Mean Agreement Ratings: "The character was probably controlled by an artificial intelligence."	119

List of Appendices

Appendix A - Virtual Agent Scripts	131
A.1 Robotic Agent Script.....	131
A.2 Humanoid Agent Script.....	140
Appendix B - Questionnaire Items	149
B.1 Pre-test Questionnaire Items	149
B.2 Empathy Questionnaire Items	150
B.3 NARS Questionnaire Items	151
B.4 Post-test Questionnaire Items.....	152
Appendix C - Protocol Forms.....	155
C.1 Informed Consent Form	155
C.2 Debriefing Form.....	160
C.3 SONA Posting.....	162
Appendix D - “Twenty Questions” Item Descriptions	163
D.1 Inanimate Objects:	163
D.2 Animate Objects:.....	163
Appendix E - Experimenter’s Reference Materials.....	164
Appendix F - Ethical Clearance.....	166

Introduction

Turing tests may be understood as human-machine interactions that are importantly similar to Alan Turing's (1950) original Turing test – the “Imitation Game,” as he called it. Since Turing, Turing tests have largely been viewed as good for one of two broad things: for prompting philosophical reflection on the limits of our concept of the mental, or for addressing empirical questions about the machines involved in human-machine interactions. In this investigation, I shall argue for a novel understanding of the usefulness of Turing tests. Specifically, I argue that, although Turing tests are certainly useful in these two ways, they are also useful in another way, which has not (to my knowledge) been proposed elsewhere in the literature. My novel proposal is that Turing tests are also good for addressing empirical questions about the *humans* involved in human-machine interactions, i.e. they are useful as what I shall call “reflexive experimental apparatus.”

I lay the foundations for this novel proposal in Chapter One, which concerns the two received ways of thinking about the usefulness of Turing tests that I mentioned above. I begin by describing the Imitation Game and noting the relevant sense in which Turing tests are importantly similar to it despite differing from it in various ways. I also spend some discussion on clarifying what Turing means by “machine” and “to think,” for he proposed the Imitation Game in order to answer the question “Can machines think?” I show that Turing's understanding of “machine” is limited to a special kind of discrete-state machine, i.e. the digital computer. By drawing from some of Turing's other work, and from a number of other thinkers' discussion of said work, I point out that in spite of

his reluctance to define what it is to think, Turing's understanding of thinking, or of the mental, is Cartesian in spirit. That is, for Turing, whether an entity can be said to think depends upon whether or not it possesses the capacity for domain-general reasoning, the capacity to express its reasoning using language, and the capacity to reason "originally," i.e. in a way that is appropriate to the situation and yet not determined by other reasoning agents. Along the way, I shall consider some of Turing's other important contributions to the cognitive sciences that will supply some motivational force for the novel understanding of the usefulness of Turing tests that I provide herein. Of importance here are the concepts he introduced that made possible the computational-representational theory of mind, i.e. computationalism. For on computationalism, the question of whether machines can think – the questions Turing answers with the Imitation Game – is an easy question to answer.

Once I have prepared the foundation for my discussion in this way, I go on in Chapter One to highlight the two understandings of the usefulness of Turing tests that I mentioned above – as "conceptual prompts" (i.e. as imaginary scenarios designed to prompt philosophical reflection on the limits of our concept of the mental) or as what I call "non-reflexive experimental apparatus" (i.e. as experimental tools for addressing empirical questions about the machines involved in human-machine interactions). Once I've illustrated these two understandings of the usefulness of Turing tests, I provide a sketch of my novel understanding of the usefulness of Turing tests as reflexive experimental apparatus. I propose that Turing tests are useful for addressing (at least) the following two sorts of empirical questions about the humans involved in human-machine interactions:

(a) Are there variable characteristics of people that play some role in their judgements of whether machines are capable of thinking, possess mental states, emotions, and the like?

(b) Do people's judgements about these things also depend in some way upon how the machines look and behave?

In other words, I will argue that, when deployed as reflexive experimental apparatus, Turing tests can help cognitive scientists to better understand aspects of the very thing they are called upon to study: the human mind.

In Chapter Two, I make the case that Turing tests can be usefully deployed as reflexive experimental apparatus by discussing both the sorts of machines, and the sorts of empirical questions about people, that I see such tests involving. By drawing upon the work of Steven Harnad, I argue that although the machines involved in Turing tests may be situated at various levels of indistinguishability from human agents, Turing tests can be fruitfully deployed as reflexive experimental apparatus even when the machines they involve are situated at relatively low levels of indistinguishability from human beings. I then proceed to supply an analogy directed at clarifying this proposal by drawing upon the work of android scientist Hiroshi Ishiguro. Just as he has used Turing tests to address empirical questions about the machines involved in human-machine interactions, where those machines are situated at lower levels of indistinguishability, I think that we can also make use of Turing tests to address empirical questions about the humans involved in such interactions, even when the machines are situated at relatively low levels of indistinguishability.

I then proceed to discuss the sorts of empirical questions about the humans involved in human-machine interactions that I think Turing tests, deployed as reflexive experimental apparatus, can help researchers to address. As I shall illustrate, there is a great deal of interest in empathy's role in human-machine interactions, especially in human-robot interaction (HRI) research. I therefore think that a good starting place is to consider empirical questions that concern empathy. More specifically, these research questions will concern the capacity of people to empathize with, and to perceive mental attributes of, the machines involved in human-machine interaction.

In Chapter Three, I present an exploratory study that serves as a “proof-of-concept” of my novel understanding of the usefulness of Turing tests. That is, I deploy a Turing test as reflexive experimental apparatus to answer questions of sorts **(a)** and **(b)**. Specifically, this study employs a virtual reality (VR) apparatus to address questions about the role that people's levels of dispositional empathy play in the judgements they make about what kind of reasoning agent – a human, or a computer program – is in control of the machines with which they interact in a virtual environment. I focus primarily on a very interesting observation from this study, namely: participants with high levels of dispositional cognitive empathy made different judgements than did those with low levels of dispositional cognitive empathy with respect to the question of whether an artificial intelligence was in control of the machine with which they interacted. Specifically, participants with higher levels of dispositional cognitive empathy judged that the machines with which they interacted were controlled by an artificial intelligence with greater certainty than did those who possess lower levels of dispositional cognitive empathy.

I suggest that these findings might be explained by virtue of the fact that those who possess high dispositional cognitive empathy are (1) good at working out what kinds of reasoning agents they are interacting with owing to high dispositional mindreading abilities, and (2) are very familiar with examples of artificial intelligence that behave similar to the virtual machines used in this study, e.g. Apple's Siri, Amazon's Alexis, Google Home, or Microsoft's Cortana. Conversely, those who possess low dispositional cognitive empathy, while also familiar with such artificial agents, might possess lower dispositional mindreading abilities.

I end by concluding that, while a certain explanation of these results is yet to be achieved, I have nonetheless identified avenues of potential research that pertains to the mindreading capacities of people, and the role this might play in human-machine interaction. Thus, I further (and more fundamentally) conclude that my proof-of-concept of the usefulness of Turing tests as reflexive experimental apparatus is successful. For not only did it achieve some interesting findings with respect to questions of sorts **(a)** and **(b)** – but especially **(a)**-type questions – further research using Turing tests as reflexive experimental apparatus can, I think, shed additional light on these findings.

So, let us begin this investigation by establishing a proper foundation from which to proceed, by discussing the ways in which the usefulness of Turing tests has been understood in the relevant literature since Alan Turing first proposed the Imitation Game.

Chapter One: On What Turing Tests Are Good For

“Can machines think?” is the question that Alan Turing set out to answer in his 1950 essay, “Computing Machinery and Intelligence.” Few would disagree that Turing’s answer to this question has had a monumental influence upon discussion concerning the nature of the mind within the cognitive sciences. Arguably, Turing himself didn’t foresee the scale of his influence on that discussion. But perhaps he would be delighted, were he alive today, to see the liveliness of that discussion. Perhaps he would even be pleased about the present investigation, which makes a novel contribution to that discussion.

This chapter concerns ways of thinking about the usefulness of Turing tests – human-machine interactions that are importantly similar to “the Imitation Game” that Turing (1950) famously proposed (and is nowadays often simply called “the Turing Test”). In section 1.1, I describe the Imitation Game, drawing from Turing’s original paper, and from some of his later work as well. I then note the relevant sense in which Turing tests are importantly similar to the Imitation Game despite differing from it in various ways. In section 1.2, I clarify what exactly Turing means by “machine,” and by “to think.” I show that Turing’s understanding of thinking machines is straightforwardly limited to a special kind of discrete state machine, namely: the digital computer.¹ I also argue, following others, that Turing’s understanding of thinking is Cartesian in flavor. I point out that for Turing, whether an entity can be said to think turns upon the following: the capacity for domain-general reasoning; the capacity to express one’s reasoning, or the

¹ My own talk of “machines” throughout will accordingly be limited to digital computers and robots (understood as embodied digital computers), for they are the kinds of machines Turing had in mind.

conclusions of one's reasoning, using language; and, the capacity to do so "originally," i.e. in a way that is appropriate to the situation and yet not determined by other reasoning agents. I also point out some of Turing's other contributions to the cognitive sciences, particularly the concepts he introduced which made possible what is now known as the computational-representational theory of mind, or computationalism. On computationalism, the question of whether machines can think is quite easy to answer. As I discuss below, the computationalist answer to the question "Can machines think?" supplies motivational force for a new way of thinking about what Turing tests are good for, namely: that they are good for addressing empirical questions about the minds of the humans involved in human-machine interactions.

Before I illustrate this new way of thinking about what Turing tests are good for, I shall discuss the two principal ways in which the usefulness of Turing tests has been understood to date. I do so after having laid out what Turing means when he uses terms like "machine" or "to think." The first way in which the usefulness of Turing tests has been thought of to date has seen them as good for prompting philosophical reflection on the limits of our concept of the mental (e.g. because they are good at eliciting intuitions about the proper application of this concept). I call this way of thinking about what Turing tests are good for the view according to which Turing tests are useful as "conceptual prompts,"² and it is the focus of section 1.3. The second way in which the usefulness of Turing tests has been conceived to date, discussed in section 1.4, is one according to which they are useful as what I call "non-reflexive experimental apparatus":

² When discussing the usefulness of the Turing tests as conceptual prompts, many refer to it as a thought experiment (cf. the collected writings in Shieber 2004). What I'm calling "conceptual prompts" here are thought experiments in the philosophical sense, i.e. imaginary scenarios designed to elicit our intuitions about how to properly apply concepts.

as experimental tools for addressing empirical questions about the machines involved in human-machine interactions (e.g. by determining whether the verbal behavior of the machines is sufficiently indistinguishable from that of humans). Turing hints at both of these ways of thinking about the usefulness of Turing tests in his own discussion of the original Imitation Game. Illustrating them here will help to set the stage for my novel way of understanding the usefulness of Turing tests.

In section 1.5 I provide a preliminary sketch of this novel understanding, according to which Turing tests are useful as “reflexive experimental apparatus”: that is, Turing tests are good for addressing empirical questions about the *humans* involved in human-machine interactions. Viewed as reflexive experimental apparatus in this sense, I will argue, Turing tests can help cognitive scientists better understand aspects of the very thing they are called upon to study scientifically.

1.1 The Imitation Game

The question Turing poses at the outset of his 1950 paper – the “Big Question,” as Stuart Shieber (2004) calls it – can seem like a difficult question to answer. That is not because it is difficult to conceive of an answer *per se*; rather, it is a difficult question to answer owing to the ambiguity that surrounds words like “machine,” “think,” “intelligence,” and so forth. Turing was acutely aware of this issue, remarking that this question is “too meaningless to deserve discussion” (1950, p. 442). But Turing must not have thought this ambiguity rendered the Big Question too meaningless, for he does spend a great deal of effort defining precisely what he means by “machine.” Specifically, Turing had in mind a very special kind of machine when he posed the Big Question, namely: the digital

computer. Moreover, there are clues that are found throughout Turing's other writings and lectures which allow one to gain a clearer sense of just what Turing means by "to think," as well. But I will return to how machines and thinking were understood by Turing in a moment. For now, I will examine the question with which Turing replaces the Big Question. This question concerns a game of sorts wherein a digital computer must convincingly answer questions put to it by a human in a human-like way. Turing (1950) calls this the "Imitation Game."

The Imitation Game is played as follows: there are three players, a man (A), a woman (B), and an interrogator (C) of either gender. C is placed in a separate room from A and B (who C knows as X and Y) and must ask each of them questions, with the aim of identifying which of the other two, A and B, is the man and the woman. In the game, C cannot see or hear A or B. Instead, C asks questions via typewritten messages, or through an intermediary. C also receives answers to the questions he or she poses the same way. A's job is to try to fool C into making the wrong guess, so he might give C misleading answers to that effect. B, on the other hand, is meant to help C make the right guess, so she provides truthful answers to C. C is successful if he or she correctly identifies that A is the man and B is the woman.

The question with which Turing (1950) replaces the Big Question is: "what will happen when a machine takes the place of A in this game?" (p. 434). Elsewhere he formulates this question even more precisely, as: "Are there imaginable digital computers which would do well in the imitation game?" (p. 442). In other words, is it conceivable that a digital computer, suitably programmed, could convince C that it is not just a human being, but a human man, and not a computer? If there are imaginable machines that could

do so, then according to Turing, one can legitimately conclude that such a computer can think. Indeed, Turing believed that there were no theoretical reasons why such a machine could not be programmed to succeed at playing the Imitation Game, or “pass the Turing test” as is often said. Moreover, he believed it was only a matter of time until digital computers possessed enough memory and processing power to play the Imitation Game well.³ For example, in his seminal paper he writes that:

I believe that in about fifty years time it will be possible to programme computers with a storage capacity of about 10^9 to make them play the imitation game so well that an average interrogator [C, in the above description of the Imitation Game] will not have more than 70 per cent chance of making the right identification after five minutes of questioning (1950, 442).

In the above-quoted passages one can find hints of two ways of thinking about usefulness of the Imitation Game itself and hence, by extrapolation, of any relevantly similar sort of human-machine interaction – any relevantly similar human-machine interaction – which I will call “a Turing test”. By suggesting that we replace the question of whether machines can think with the question of whether there are imaginable digital computers that could perform well in the Imitation Game, Turing was in effect suggesting that this game, as well as Turing tests in general, be viewed as conceptual prompts. And by suggesting that the eventual (“in about fifty years’ time”) deployment of the game would be one in which the machine was quite capable of fooling the human participants, Turing was further

³ By “memory,” Turing (1950) refers to the storage capacity of digital computers. In modern computer terminology, memory refers to RAM (Random Access Memory), which is different than storage capacity.

suggesting that the game, and Turing tests in general, be viewed as non-reflexive experimental apparatus.

Here, I use “Turing test” to refer to intentionally effected human-machine interaction scenarios wherein the human is expected to interact with the machines despite not knowing what kind of reasoning agent is in control of the machine’s actions (e.g. another human versus an artificial intelligence). Understood this way, Turing tests can differ in various ways from *the* Turing test, i.e. the Imitation Game. For example, Turing tests might involve fewer than the three players Turing (1950) initially proposed. In fact, he mentions such a variant of the game elsewhere based upon a *viva voce* examination taking place between two individuals, where the machine “has to try and pretend to be a man,⁴ by answering questions put to it, and it will only pass if the pretense is reasonably convincing” (Newman, Turing, Jefferson & Braithwaite 1952, quoted in Shieber 2004, p. 118). Turing tests may also not require that there be a clear “winner” of the Imitation Game. In other words, passing or failing the Turing test may not be as straightforward as Turing imagines. For instance, as I discuss in Chapter Two (see section 2.2) android scientist Hiroshi Ishiguro suggests that it is possible that a machine can pass the Turing test at one level of cognitive processing and fail it at another. For now, however, my point is that while Turing tests, as I understand them, can differ from the Imitation Game in various ways, they are nonetheless importantly similar to it in the sense that they are all, generally speaking, intentionally effected human-machine interaction scenarios,

⁴ I interpret this a bit more modernly. In the *viva voce* version, the computer must simply pretend to be a person. This is not just an attempt to use more inclusive language. Rather, dropping the requirement that the interrogator guess the other players’ respective genders is yet another way in which Turing tests, understood as I have described them above, can differ from the Imitation Game.

where the human must interact with the machine despite initially not knowing whether the machine's actions are controlled by another human, or by a computer program.

1.2 On Thinking Machines

Clarifying the two above-mentioned ways of thinking about the usefulness of Turing tests – their usefulness as conceptual prompts and as their usefulness as non-reflexive experimental apparatus – will be made easier if a little time is spent examining in greater detail just what Turing means when he employs words like “machine,” “intelligence,” and “to think.”

As I noted earlier, Turing (1950) spends much effort on clarifying exactly what kind of machine he has in mind when he asks the Big Question, namely: the digital computer. But he does not spend nearly as much effort defining what it is to think. Nonetheless, I think that there are important insights into what exactly Turing means by “to think” or “thinking” that can be gained from some of his other writings, lectures and interviews, besides his 1950 paper. So, before turning to Turing's understanding of digital computers, I will survey some of his important insights concerning the mental, with the aim of showing that even though Turing never explicitly defined what it is to think, he nonetheless seems to have had a very particular conception of thinking in mind.

In many subsequent discussions and analyses of his work, Turing's understanding of what it means to think – reflected in his formulation of the Imitation Game – is viewed by a number of cognitive scientists as somewhat Cartesian. For example, Noam Chomsky (quoted in Epstein, Roberts and Beber 2008), remarks that “Turing's imitation game is reminiscent of ideas that were discussed and pursued during what we might call that ‘first

cognitive revolution of the seventeenth century,” (p. 105), during which time René Descartes was a leading, if not sometimes controversial, intellectual figure. Similarly, Daniel Dennett (quoted in Shieber 2004) muses that “Perhaps [Turing] was inspired by Descartes, who in his *Discourse on Method* plausibly argued that there was no more demanding test of human mentality than the capacity to hold an intelligent conversation” (p 271). Further, Stevan Harnad (1991; 1994; 2000) has remarked that the Turing Test (i.e. the original Imitation Game) is quite closely related to the classic philosophical problem of other minds, which concerns the grounds upon which one is justified in believing that other people have minds. For Harnad, the sort of behaviour displayed by a machine that “wins” the Imitation Game is just the sort of thing that many philosophers who address the problem of other minds, including Descartes himself, have pointed to as an essential element of the solution to said problem.

Indeed, as philosophical antecedents to the Imitation Game go, the work of René Descartes is probably the most relevant to this investigation, so let us now gain a better sense of the Cartesian spirit of the game. Descartes wrote in his *Discourses on the Method* on the distinguishing characteristic of man and animals:

[I]f there were machines bearing the image of our bodies, and capable of imitating our actions as far as it is morally possible, there would still remain two most certain tests whereby to know that they were not therefore really men. Of these the first is that they could never use words or other signs arranged in such a manner as is competent to us in order to declare our thoughts to others [...] The second test is, that although such machines might execute many things with equal or perhaps greater perfection than any of us, they would, without doubt, fail in

certain others from which it could be discovered that they did not act from knowledge, but solely from the disposition of their organs; for while reason is an universal instrument that is alike available on every occasion, these organs, on the contrary, need a particular arrangement for each particular action (Descartes, quoted in Shieber 2004, pp. 27-28).

So, according to Descartes, the essential difference between human beings on the one hand, and machines on the other, is the human ability to use language to “declare our thoughts to others,” and to employ reason universally and not deterministically or mechanistically, as machines do. This is thanks to what Descartes believed separated humanity from the rest of nature and artifice: the rational soul, which he believed to be endowed in each of us by God.⁵ Descartes continues:

[B]y means of these two tests we may likewise know the difference between men and brutes. For it is highly deserving of remark, that there are no men so dull and stupid, not even idiots, as to be incapable of joining together different words, and thereby constructing a declaration by which to make their thoughts understood; and that on the other hand, there is no other animal, however perfect or happily circumstanced, which can do the like (Descartes, quoted in Shieber 2004, p. 28).

And so, using language to declare one’s thoughts to other people, and using reasoning in a universal fashion – or what cognitive scientists today would describe as domain general

⁵ Here is one point of departure between Descartes and Turing: while Descartes believed the possession of a rational soul is a key difference between humans on the one hand, and machines on the other, Turing is not committed to its existence and sees no reason in principle why machines cannot be said to think.

reasoning – are the hallmarks of intelligence for Descartes. I think that Shieber (2004) puts it quite well when he points out that it was Descartes who first “pinpointed *verbal behavior* [sic] as the crucial property for distinguishing humans from beasts, the soul-bearing from the soul-less” (p. 4).⁶ He continues, remarking that Turing’s Imitation Game is a “game-theoretic crystallization of Descartes’s observation that flexibility of verbal behaviour is the hallmark of humanness” (Ibid., p. 7).

Setting examples from the philosophical tradition aside for the moment, when we use the verb “to think” in everyday speech, we typically mean that we’re deliberating, considering something, or employing our powers of reason to solve a problem or answer a question. So perhaps the simplest way to describe what it is to think, is to define thinking as the use of reason in a very general sense. Arguably this accords very well with what Descartes thought about intelligence. But what is it to reason? This is a problem that philosophers have long grappled with, and which they – along with cognitive scientists – still grapple with today. Unlike Descartes, I don’t think positing the existence of the rational soul is necessary to explain what reason is or how it works. Instead, I think a definition of reasoning that should suffice for this investigation is a computationalist one: reasoning, at the very least, consists in transitioning from one representational state to another.⁷ These representational states might be propositional attitudes like beliefs and desires, or various symbolic representations realized within the circuitry of a digital computer. Turing seems to have this kind of thing in mind when he writes of digital computers being capable of thinking, or of displaying human-like

⁶ Emphasis is Shieber’s.

⁷ I discuss computationalism in greater detail below.

intelligence. This is not surprising since Turing is arguably as much the forefather of computationalism as he is the forefather of artificial intelligence. Furthermore, an understanding of “thinking” as the domain-general use of reason is nicely captured by Turing’s descriptions of the kinds of problems that thinking machines can solve. For example:

[I]t has been shown that there are machines theoretically possible which will do something very close to thinking. They will, for instance, test the validity of a formal proof in the system of *Principia Mathematica*, or even tell of a formula of that system whether it is provable or disprovable (Turing 1951a, quoted in Shieber 2004, p. 105).

The ability to apply reason toward solving problems is certainly commensurable with how we typically understand the verb “to think.” But Turing did not simply argue that for a machine to think, it must be able to employ reason in a human-like way. Rather, machines ought to be able to reason about a variety of things. Specifically, they ought to display what cognitive scientists today call domain-general reasoning, or what Herbert Simon and Allen Newell call “general intelligent action” (Newell & Simon 1976, p. 116). For example, consider some of the “specimen questions and answers” that Turing supplies in his paper concerning the Imitation Game. Note, “**Q**” refers to the questions supplied by the interrogator, and “**A**” refers to the answers supplied by the machine:

Q: Please write me a sonnet on the subject of the Fourth Bridge.

A: Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.

A: (Pause about 30 seconds and then give as answer) 105621.

Q: Do you play chess?

A: Yes.

Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1.

It is your move. What do you play?

A: (After a pause of 15 seconds) R-R8 mate (pp. 434-435).

I think that this example shows that the domain-general use of reasoning is what Turing has in mind when he discusses what it means for an agent to think. And of course, the verbal interaction between humans and machines in the Imitation Game highlights the role of language in declaring our thoughts to others, to borrow Descartes's terminology. I therefore contend that although Turing claimed to have thought that attempting to ascertain what it means to think is a fruitless endeavor, a simple comparison to Descartes's own test of human-like intelligence reveals two important insights into what Turing believed constitutes thought, or the mental: (1) the ability to use reason in a domain-general way; and (2), to express one's reasoning, or the conclusions of one's reasoning, using language.

But that is not the whole story. Turing (1950) also names one other Cartesian criterion for thinking machines to count as such: freedom of the will. Consider, for example, one of the objections to the prospect of thinking machines that Turing calls Lady Lovelace's Objection:

[Scientists'] outlook was well summed up by Lady Lovelace over a hundred years ago, speaking of Babbage's Analytical Engine. She said [...] "The Analytical

Engine has no pretensions whatever to *originate* anything. It can do whatever *we know how to order it to perform.*” This very well describes the way in which digital computers are actually used at the present time, and in which they will probably mainly be used for many years to come. For any one calculation the whole procedure that the machine is to go through is planned out in advance by a mathematician. The less doubt there is about what is going to happen the better the mathematician is pleased. (Turing 1951b, quoted in Shieber 2004, p.111-12).

That is to say that while human verbal behaviour is apparently very flexible, the various behaviours of a computer are not. As Turing (1951b) himself put it “To behave like a brain seems to involve free will, but the behaviour of a digital computer, when it has been programmed, is completely determined” (Quoted in Shieber 2004, p. 114). More specifically, the digital computer’s behaviour is completely determined by the program, which is of course determined by a programmer. The deterministic responses of any computer might easily give it away as such in a Turing Test, and so Turing concluded “that a machine which is to imitate a brain must appear to behave as if it had free will” (Ibid.). As a solution, Turing proposes that a “random element” be incorporated into any computer that is subjected to the Turing Test:

This would result in the behaviour of the machine not being by any means completely determined by the experiences to which it was subjected, and would have some valuable uses when one was experimenting with it (Turing 1951a, quoted in Shieber 2004, pp. 108-109).

Arguably, Descartes himself would have agreed with Turing on this point, for he believed that the behaviour of animals – which, as mentioned, he surmised were simply intricate machines designed by God – was also completely determined.⁸ Turing and Descartes would probably also agree that while verbal behaviour must not be completely determined, it also cannot be completely random. Rather, the verbal behaviour of any thinking agent – be it a human or a machine – must be appropriate to the situation. Thus, even though Turing spends very little effort on explicitly defining what it is to think, we can nonetheless establish that Turing must have held that the following were constitutive of what it is to think:

- (1) The ability to use reason in a domain-general way.
- (2) The ability to express one's reasoning, or the conclusions of one's reasoning, using language.

And to these two criteria, one might add something like the following:

- (3) The capacity for original reasoning.

By (3) I mean that the machine's actions – including its verbal behaviour – appear not to be completely determined by antecedent causes that originate from other, thinking agents, such as the machine's programmers, or its interrogators. For Turing seems to be especially concerned with the machine's behaviour being determined by its program, which is ultimately itself determined by another agent who is capable of (1)-(3), i.e. the machine's programmer. But the behaviour of the machine, if it to pass the test, must not

⁸ As mentioned, unlike Descartes, Turing does not commit himself to the idea of the immaterial, rational soul.

appear to be completely determined by either of these antecedent causes. It must also not appear to be determined by the inputs it receives from other reasoning agents, i.e. what people say to the machine in a Turing Test. And yet, the machine must also respond appropriately to what is said to it, lest its behaviour appear so random as to seem stupid and mindless. But what matters where **(3)** is concerned is that the machine's behaviour appears to originate from the machine itself, and not from some other reasoning agent.

I say that the machine must *appear* to function this way because it is certainly possible that the machine's behaviour may in fact be determined by some other antecedent causes besides other reasoning agents. It is entirely possible, however, that even though we humans have the sense that our thoughts and actions originate from within us, they may also be determined. But I will leave this question open, since this is not an investigation into whether we have freedom of the will.

If **(1)**, **(2)** and **(3)** in fact capture what Turing took to count as "thinking," and further, if the Imitation Game is a reliable means of determining whether a machine has characteristics **(1)**, **(2)** and **(3)**, then the Imitation Game ought to serve as a good test of whether an agent – artificial or natural – can be said to think. I consider this further in the following sections of this chapter. But philosophical antecedents and exegetical considerations about Turing's body of work aside, I have shown that many agree that Turing's understanding of what it is to think is Cartesian in spirit. Where Turing and Descartes differ seems to primarily concern the nature of the mind: Descartes believed in the immaterial, rational soul; Turing, a physicalist, saw no reason why a machine such as a computer could not think like people do. But such metaphysical discussions are best saved for another occasion.

So far, I have made some headway at making sense of what Turing means by “to think” and thus at making sense of his conception of the mind. But Turing made another important contribution to the cognitive sciences which may help us to side-step some of the tricky philosophical questions that concern the nature of thought altogether. I consider these contributions below, for they can also provide some motivation for thinking of the usefulness of Turing tests along the novel lines I propose, i.e. for thinking of Turing tests as reflexive experimental apparatus.

Probably Turing’s most important contribution to computer science and artificial intelligence is the Turing machine. Now, Turing did not posit the idea of a Turing machine in order to address the Big Question. Rather, Turing was seeking a way to address what is known as the Halting Problem – the problem of determining, given a function, a computer program written to solve that function, and a set of inputs, whether the program will identify a solution or continue running infinitely. The theoretical computing machines that Turing imagined – which we now call Turing machines – helped him to establish that there is no solution to this problem (Turing 1936).

A Turing machine is a fictional entity. It contains an infinitely long tape, so that the machine has no storage limitations. The tape contains cells, onto which symbols can be written or deleted. The symbols on the tape represent information. For example, 0 can represent “false” while 1 might represent “true.” The machine also has a machine head, which can read the symbols on the tape, delete symbols, write symbols, and which can move the tape to the left or to the right one cell at a time. This machine also has a machine table which contains a set of instructions for how the machine is to operate. The

instructions are also represented using symbols like 0s and 1s. This is what we would nowadays describe as a computer program.

For Turing, a key feature of these machines is that they can be represented mathematically, their states as series of numbers and so forth. Simply put, this allowed Turing to derive certain mathematical proofs about the properties of these machines. One of the proofs he formulated using his imaginary machines was that there is no solution to the Halting Problem. Yet another was that if Turing machines are possible so too is a universal Turing machine, a special kind of machine upon which any specific Turing machine (for arithmetical calculations, for instance) can be implemented. The universal Turing machine is, in essence, the theoretical forerunner of the modern digital computer. But most importantly for my purposes, it was in the context of discussing Turing machines and the universal Turing machine that Turing (1936) sowed the seeds of a view that is now widely accepted within the cognitive sciences: the computational-representational theory of mind, or sometimes simply “computationalism” (cf. Crane 2003; Thagard 2005; Clark 2014).

As I just mentioned, the universal Turing machine is regarded as the theoretical precursor to the modern digital computer. Like the universal Turing machine, and indeed any of the Turing machines which can run on it, modern digital computers are an example of what Turing (1950) called a “discrete state machine” (p. 439). Such machines exist in one definite state at one moment in time. According to Turing, one can think of a light switch as a very simple discrete state machine: it is either “on” or “off” (pp. 439-40). Digital computers work on the very same principle. Modern digital computers encode information in a binary language using 1s and 0s. A 1 can be likened to a switch

in the “on” position while a 0 can be likened to a switch in the “off” position. (Or “true” and “false,” as I indicate above). A digital computer can therefore be thought of as a very large collection of “on/off” or “true/false” switches. When the computer performs an operation, it changes state one step at a time, proceeding through a finite number of states as it executes its program.⁹

In other words, like Turing machines, discrete state machines such as digital computers carry and transform what those in the cognitive science call symbolic representations. The 1s and 0s that make up the binary language are represented with physical patterns, i.e. the configuration of switches or circuits that are on or off. And of course, the various programs that we might run on such a computer – the equivalent of a Turing machine’s machine table – are written in the 0s and 1s of binary. Whatever the program, it contains rules that the computer must follow as it transforms a given input into an output. As the computer runs the program, it changes state, step by step, until whatever information processing task it has been put to is complete.

Turing provided the theoretical tools that made this understanding of computation possible. Indeed, his Turing machines, together with the idea of the algorithm, i.e. a step-by-step set of rules governing how representations are to be transformed into other representations, are the very foundations of the computational-representational theory of mind. On this theory, computation is just the rule-governed transformation of representations (symbols, i.e. physical patterns). Furthermore, thinking, as the core aspect of the mental or the cognitive, is just a kind of computation. The mind works much the

⁹ The idea that symbols – which are physical patterns – can be used to represent information is a key aspect of the Physical Symbol System Hypothesis (cf. Newell and Simon 1976), to which I return below.

same way as a computer – indeed, it is a natural computer – in that it carries out the very same kinds of rule-governed, algorithmic symbol manipulation as does a Turing machine or a digital computer. Note further that the definition of reasoning I gave earlier accords with the understanding of cognition that I have just described.

Let's now return to the Big Question. Recall that Turing intended for the Imitation Game to serve as a replacement to this question. But if one is a computationalist – as many cognitive scientists are – then one need not replace the Big Question in the first place. To the computationalist, “Can machines think?” is not an empirical question. Rather, it is a theoretical question with a simple answer. Yes, machines like digital computers can think, because cognition and computation are ultimately the same thing: the rule-governed transformation of representations that carry information. I cannot help but think it is somewhat ironic that Turing (1950) goes to such great lengths to replace the Big Question with the Imitation Game. As I have shown, he thought that this was a more tractable form of the Big Question. However, prior to this he introduced the conceptual tools that allowed for perhaps the most widely-accepted answer to the Big Question to be realized. As I show in section 1.5, this insight puts some additional motivational force behind my novel way of thinking about the usefulness of Turing tests.

Of course, not all cognitive scientists are computationalists. And of course, the computationalist answer to the Big Question is not the only answer to the Big Question. If by “thinking” one simply means computation as the rule-governed transformation of symbols, that is fine, but this is undoubtedly going to prove an unsatisfactory answer to some. For there are those that deny that computational information processing systems – even if they possess **(1)** the ability to use reason in a domain-general way, **(2)** the ability

to express that reasoning using language, and **(3)** the capacity for original reasoning – are intelligent (see below). I will not try to decide this matter here. However, now that I’ve sketched a clear picture of the kinds of machines that Turing would have play the Imitation Game, and a clearer understanding of what Turing means by “to think,” I can proceed with laying the groundwork for the remainder of this investigation.

1.3 Turing Tests as Conceptual Prompts

As I pointed out at the start of this chapter, the reason Turing thought that the question “Can machines think?” is meaningless and therefore undeserving of discussion owes to the ambiguity surrounding the words “think” and “machine.” However, as I have shown, Turing spends a great deal of effort specifying exactly what kinds of machine he has in mind when it comes to the Imitation Game: the digital computer. I also argued that Turing’s suggestion that it is meaningless to ask whether machines can think is somewhat ironic. Turing was, after all, not the first person to conceive of a digital computer. But he was the first person to try to formalize the notion of computation (cf. Turing 1936). To do this he used an imaginary kind of digital computer that we now call a Turing machine (explained above), laying the foundations of computationalism.

Let us now reorient this discussion to one of the two principal ways in which the usefulness of Turing tests has been conceived: as conceptual prompts, or as prompting philosophical reflection on the limits of our concept of the mental. Recall (from 1.1) that by suggesting that we replace the question of whether machines can think with the question of whether there are imaginable digital computers that could perform well in the Imitation Game, Turing was in effect suggesting that Turing tests can be viewed as

conceptual prompts. Owing to technological limitations, it was not possible to build such digital computers during Turing's lifetime, but that doesn't matter here. The point is that, for Turing, there was no reason *in principle* why a digital computer, with enough memory, storage, and processing power, could not be programmed to do well at the Imitation Game, or at Turing tests in general. And since the machine would be displaying **(1)** an ability to reason in a domain-general way, **(2)** an ability to express said reasoning using language, and **(3)** a capacity for original reasoning by virtue of playing the game well, we would in Turing's view have good reason to say that it is possible for the machine "to think," i.e. we would have good reason to extend our concept of the mental to artifacts like it.

Turing himself, then, seems at least to have suggested that Turing tests can be useful for prompting philosophical reflection on the limits of our concept of the mental. Another illustration of how the usefulness of Turing tests has been viewed along these lines – as conceptual prompts – can be drawn from John Searle's (1980) famous "Chinese Room" thought experiment. A staple of philosophy and cognitive science textbooks, Searle begins by having the reader imagine a computer program that can receive inputs, which in this case are whole sentences which Searle calls "stories," and answer questions put to it about the stories. He then asks, if the computer were able to answer these questions in a Turing test-style question and answer session, then would we be justified in claiming the computer *understands* the questions being put to it, or the answers it generates? Searle thinks that the answer is "no," for, he argues, the computer is blind to the meaning – the semantics – of the symbols that make up its program, the inputs it receives, the outputs it generates, *et cetera*. Indeed, the computer computes – i.e.

transforms symbols in a rule-governed, algorithmic way – by operating only on the formal properties, i.e. the syntax, of its symbols.¹⁰

To clarify his negative answer here, Searle suggests that we replace the computer in his thought experiment with a human being, having the human run the program in the computer's stead. He thus has us imagine that this person is situated within a room, and that the stories she receives, and the responses she generates, are passed in and out of the room on sheets of paper. Searle further has us imagine that the symbols being passed in and out of the room are Chinese characters, thus the "Chinese Room."

Here, the human who replaces the computer is given sheets of paper with Chinese characters on them (I assume Mandarin Chinese here), and she must send an appropriate response to these inputs back out of the room to the "operators" of the computer, who are native Chinese speakers. The person inside of the room doesn't speak Chinese, but she has a set of instructions written in English which outline the appropriate responses to the strings of symbols that the operators pass into the room. She must use her set of instructions to identify what response to deliver, based upon her book of instructions and upon which inputs she receives. But like the computer mentioned above, she has no access to the semantic properties of the symbols that she operates on, for she doesn't understand Chinese writing. Further, her set of English instructions is not a means of translating Mandarin into English, then back to Mandarin. Rather, it is merely a set of instructions written in English that outlines all of the rules for responding to the Chinese

¹⁰ "Formal" refers to the syntactic properties, or the physical characteristics, of the symbols. Note further that some readers will find Newell's and Simon's (1976) discussion of the Physical Symbol System Hypothesis helpful for providing some additional context to Searle's objections about whether it is possible for an AI to possess understanding.

characters according to their formal or syntactic properties, not their semantic properties (cf. Searle 1980, especially pp. 417-19).

Except for the slow processing time when compared to a digital computer, Searle argues that the human computer inside of the Chinese Room would be indistinguishable from a digital computer, or a native speaker of Mandarin inside of the room. In other words, the person in the Chinese Room passes the Turing test with respect to understanding Mandarin. Nonetheless, Searle argues, it is very intuitive that the digital computer would have no more understanding of the questions put to it (or of the answers it generates) than the person who replaces it in the thought experiment has understanding of the Mandarin characters she receives and transmits, which is to say no understanding at all. Hence, on the basis of his thought experiment, Searle suggests that we should *not* extend our concept of the mental to digital computers as Turing thought we perhaps should – that, contrary to Turing’s suggestion, we should not see mental concepts like “understands” and “thinks” as properly applicable to machines like digital computers.

It is not the aim of this chapter to adjudicate between these two perspectives, Searle’s or Turing’s, on the philosophical question of whether the limits of our concept of the mental extend far enough to comprehend digital computers. For those who are interested, many relevant writings on this sort of philosophical debate – including Searle’s and Turing’s perspectives – are collected in Shieber (2004). My point in mentioning Searle’s perspective is rather to emphasize that it too, just like Turing’s conflicting perspective, takes Turing tests, in the sense laid out above, to be useful as conceptual prompts – again, as things that prompt us to consider philosophical questions about the limits of our concept of the mental. For notice that Searle’s entire thought

experiment begins with a Turing test, in the sense I have laid out: before being replaced by the human being in the thought experiment, after all, the digital computer is involved in a Turing test in that sense.

1.4 Turing Tests as Non-reflexive Experimental Apparatus

In 1966 Joseph Weizenbaum, a computer scientist at MIT, published a noteworthy paper in which he described a computer program called ELIZA. ELIZA was the precursor to what we now call “chatbots,” i.e. computer programs where sentences are typed into a terminal, and the chatbot returns a human-like response. Such computer programs are meant to simulate the natural conversations that occur between people, and they are an example of the kind of AIs that are most often involved in recent Turing tests.

ELIZA was programmed by Weizenbaum to converse in the style of a Rogerian psychotherapist, i.e. a style of psychotherapy developed by the humanist psychologist Carl Rogers. Without delving too deep into the technical details, using only about 200 lines of code ELIZA was able to search for keywords in the sentences that it received, and use those keywords – along with a number of “canned” responses – to generate a reply. An oft-cited example of a conversation with ELIZA – between ELIZA and Weizenbaum’s secretary to be precise – is as follows:

Human: Men are all alike.

ELIZA: In what way?

Human: They’re always bugging us about something or other.

ELIZA: Can you think of a specific example?

Human: Well, my boyfriend made me come here.

ELIZA: Your boyfriend made you come here?

Human: He says I'm depressed much of the time.

ELIZA: I am sorry to hear that you are depressed.

Human: It's true. I am unhappy.

ELIZA: Do you think coming here will help you not to be unhappy?

Human: I need some help, that much seems certain.

ELIZA: What would it mean to you if you got some help?

Human: Perhaps I could learn to get along with my mother.

ELIZA: Tell me more about your family.

Human: My mother takes care of me.

ELIZA: Who else in your family takes care of you?

Human: My father.

ELIZA: Your father?

Human: You are like my father in some ways.

ELIZA: What resemblance do you see?

Human: You are not very aggressive but I think you don't want me to notice that.

ELIZA: What makes you think I am not very aggressive?

Human: You don't argue with me.

ELIZA: Why do you think I don't argue with you?

Human: You are afraid of me.

ELIZA: Does it please you to believe I am afraid of you?

Human: My father is afraid of everybody.

ELIZA: What else comes to mind when you think of your father?

Human: Bullies.

ELIZA: Does that have anything to do with the fact that your boyfriend made you come here? (Adapted from Weizenbaum 1966, pp. 36-37).

Weizenbaum (1966) observed that “Some subjects have been very hard to convince that ELIZA (with its present script) is *not* human. This is a striking form of Turing’s test. What experimental design would make it more nearly rigorous and airtight?” (p. 42).¹¹

Indeed, conversational interactions like those that took place between ELIZA and people at MIT in the 1960s feature in perhaps the most famous – or infamous – events involving Turing tests: the Loebner Prize competition. Established in 1990 by Hugh Loebner, this annual competition features significant monetary prizes awarded to the programmers of the machines that most successfully fool the judges participating in the relevant Turing tests.¹²

AI researchers have voiced criticism of the Loebner Prize competition on the grounds that it involves poor experimental apparatus, and/or that it is of little value to AI research. Indeed, Loebner’s competition is exactly that: a competition. It is not aimed at answering well-formulated research questions nor testing precisely formulated hypotheses. Thus Stevan Harnad (1992) has written that Loebner’s competition is “trivial from a scientific standpoint” (pp. 9-10). Moreover, descriptions of the competitions tell of yearly disorganized, chaotic gatherings whose organizers are almost never invited back to participate at the same venue! (Sundman 2003). AI pioneer Marvin Minsky, who

¹¹ Note that when Weizenbaum talks of a “form of Turing’s test,” he appears to mean “a Turing test” in our sense of the term (see 1.1 above).

¹² See, for example, the rules and regulations of the 2018 competition: <http://aisb.org.uk/events/loebner-prize#finals2018>

offered a prize of his own to anyone who could convince Hugh Loebner to cease his competition, viewed the competition as an intellectually empty publicity stunt, saying that:

I do hope [...] that Mr. Loebner will indeed revoke his stupid prize, save himself some money, and spare us the horror of this obnoxious and unproductive annual publicity campaign. In fact, I hereby offer the \$100.00 Minsky prize to the first person who gets Loebner to do this. (Minsky, quoted in Sundman 2003).

The scientific merits (or lack thereof) of the Loebner Prize competition aside, it is perhaps the most well-known event in which Turing tests are used as tools for addressing empirical questions about the features of the machines involved in human-machine interactions – empirical questions about the degree to which the programmed computers are capable of displaying convincingly human-like verbal behavior, for example. Thus, the Loebner Prize competition serves as an illustration of the second principal way in which the usefulness of Turing tests has been conceived since Turing himself: as what I have called non-reflexive experimental apparatus.

1.5 Toward Re-Imagining What Turing Tests are Good For

So far, I have surveyed Alan Turing’s original formulation of what is now widely known as the Turing test, i.e. the Imitation Game. Specifically, in Section 1.1, I gave a comprehensive overview of the Imitation Game. In Section 1.2, I made certain that discussion about Turing tests in this investigation will remain concise by clarifying what Turing (1950) means when he uses terms like “machine” or “to think.” Along the way, I

explored an important philosophical antecedent to Turing's work, i.e. the work of René Descartes. This was done in order to illustrate the Cartesian nature of the conception of intelligence, or mentality, that Turing seems to have had, something that has been noted by a number of cognitive scientists. I argued that Turing must have held that thinking things have the following characteristics: **(1)** the ability to use reason in a domain-general way; **(2)** the ability to express one's reasoning, or the conclusions of one's reasoning, using language; and **(3)** the capacity for original reasoning. In the course of this discussion, I also illustrated how Turing's work also laid the groundwork for a view called computationalism. As I've shown, one's inclinations toward computationalism arguably entail certain positions with respect to Turing tests.

Of course, the main objective of this chapter, besides clarifying how Turing understood thinking machines, is to illustrate the two principal ways in which the usefulness of Turing tests – what they are good for – has been conceived since Turing's monumental work. In Section 1.3, I discussed the conception of Turing tests as being good for prompting philosophical questions about the limit of our concept of the mental, i.e. the conception according to which Turing tests are useful as conceptual prompts. Here is where we encountered Searle's (1980) reply to Turing's argument in the form of yet another conceptual prompt, the so-called Chinese Room. In Section 1.4, I presented the conception of Turing tests as being good for addressing empirical questions about the machines involved in human-machine interactions. Here we saw this conception of the usefulness of Turing tests vividly illustrated in the (in)famous Loebner Prize competition. Of course, my discussion of Turing's own work, and of those two principal ways in which the usefulness of Turing tests has been conceived since his work, is not meant to

be exhaustive. For those who are interested in a more thorough account of Turing's work, and of Imitation Game as it has traditionally been approached, French's (2000) paper is brief yet quite comprehensive. For even more in-depth accounts, I recommend the collections of Shieber (2004) and Epstein, Roberts and Beber (2008).

With the central objective of this chapter met, I am now in a position to make the case for a re-imagining of what Turing tests are good for – to make the case they should be thought of as useful not simply as conceptual prompts or as non-reflexive experimental apparatus, but rather as reflexive experimental apparatus. This novel way of thinking about the usefulness of Turing tests, as we will see in the following chapter, is similar to the non-reflexive experimental apparatus conception by virtue of seeing Turing tests as useful experimental tools for addressing empirical questions. Unlike the non-reflexive apparatus conception, however, and rather more like the conceptual prompt conception, the reflexive experimental apparatus conception I will defend sees Turing tests as useful addressing empirical questions about the *humans* involved in human-machine interactions, and hence ultimately about ourselves.

Two sorts of empirical questions about the humans involved in human-machine interactions that Turing tests are useful for addressing, I will suggest, are as follows:

- (a) Are there variable characteristics of people that play some role in their judgements of whether machines are capable of thinking, possess mental states, emotions, and the like?
- (b) Do people's judgements about these things also depend in some way upon how the machine looks and behaves?

And, indeed, in the final, “proof of concept” chapter of this investigation, I will lay out a preliminary attempt to show that Turing tests are useful for dealing with particular questions of both of these sorts. Before that, however, I need to say more about the kinds of machines, and the kinds of empirical questions about the people involved in human-machine interactions, that I see Turing tests employed as reflexive experimental apparatus as involving. This is the subject of the next chapter, to which we now turn.

Chapter Two: On the Usefulness of Turing Tests as Reflexive

Experimental Apparatus

At the end of the previous chapter, I introduced a new conception of the usefulness of Turing tests: as reflexive experimental apparatus, i.e. experimental tools for addressing empirical questions about the humans involved in human-machine interactions. This new conception, I argued, is distinct from the two principal conceptions of the usefulness of Turing tests since Turing's (1950) introduction of the original Turing test, the Imitation Game: as conceptual prompts (i.e. as tools for prompting philosophical reflection on the limits of our concept of the mental) or as non-reflexive experimental apparatus (i.e. as experimental tools for addressing empirical questions about the machines involved in human-machine interactions). Although I think it is true that Turing tests are useful in the ways indicated by those two principal conceptions, I think Turing tests are also be useful in another significant way, a way which has not to my knowledge been discussed (or even explicitly proposed) elsewhere in the literature.

In order to make the case that Turing tests can indeed be good for addressing empirical questions about the humans involved in human-machine interactions, I need to say more about both the sorts of machines and the sorts of empirical questions about the humans I see these tests as involving. Drawing upon the work of Steven Harnad, I will begin in section 2.1 by noting that although the machines involved in Turing tests may be situated at various levels of indistinguishability from human agents, Turing tests can be fruitfully deployed as reflexive experimental apparatus even where the machines they involve are situated at lower levels of such indistinguishability.

In section 2.2 I provide an analogy to help clarify this: just as the android scientist Hiroshi Ishiguro has made use of Turing tests to address empirical questions about the machines involved in human-machine interactions, where the machines involved are situated at relatively low levels of indistinguishability from human agents, so we can expect to make use of Turing tests to address empirical questions about the humans involved in such interactions even when the machines involved in the tests are situated at relatively low levels of indistinguishability.

I go on in section 2.3 to discuss the sorts of empirical questions about the humans involved in human-machine interactions Turing tests as reflexive experimental apparatus can fruitfully address. These sorts of empirical questions pertain to empathy, for as I shall show in a moment, empathy's role in human-machine interactions is of great interest to those who study how people engage with ostensibly thinking machines. This discussion will allow me to formulate more specific research questions than my first general research question – which concerns the extent to which variable characteristics of people play some role in their judgements of whether machines are capable of thinking, possess mental states, emotions, and the like – in the exploratory study I present in my final chapter. More specifically, these research questions will concern the capacities of empathy, and of perceiving attributes of the mental, of people involved in human-machine interaction Finally, in section 2.4, I summarize the considerations of this chapter.

2.1 Levels of Indistinguishability and Harnad's Turing Test Hierarchy

First, let us consider Stevan Harnad's Turing test hierarchy (cf. Harnad 1991; 1992; 1994; 2000). Harnad's hierarchy contains five different levels. In each of these levels, the machine involved in a Turing test is meant to be indistinguishable from human agents along a certain dimension or dimensions of indistinguishability. The first, lowest level of Harnad's Turing test hierarchy is level t1, which Harnad also calls the "toy model" level (which is why a lower-case letter T is used). Toy models, according to Harnad, capture just a fraction of the cognitive capacities that they model and are thus meant to be minimally indistinguishable from the things they model. For example, as Harnad points out, a toy boat will float just like a real one, yet under normal circumstances we would never mistake one for the other. At first pass, a computer program such as ELIZA (discussed in section 1.4) would seem to be situated at level t1. Furthermore, Harnad thinks that just about all cognitive modelling that currently takes place within the cognitive sciences occurs at this level in the hierarchy. He does, however, believe that it is in principle possible to move further up to level T2.

At level T2, the machine involved in the Turing test is meant to be indistinguishable from a human agent primarily in terms of its linguistic and computational capacities. This is the level of the hierarchy at which Turing (1950) envisions his Imitation Game. Level T2 is also the level at which the computer is situated in Searle's (1980) Chinese Room argument (see Section 1.3). Since I discussed the Turing test at length in the previous chapter, I will not spend any more discussion on Level T2 for the moment.

At level T3 in Harnad's Turing test hierarchy, the machine involved in the test is meant to be indistinguishable from a human in terms of both its linguistic capacities, and what Harnad calls its "robotic" capacities. Harnad calls this level of the hierarchy the "Total Turing Test," or sometimes called the "Robotic Turing Test." What Harnad means is that at level T3 the machine is supposed to be indistinguishable from a human in terms of not just its linguistic capacities, but also its behavioural capacities and its physical appearance. Here, the machine's physical appearance and behaviour matter just as much as its verbal behaviour, to borrow Shieber's (2004) terminology. At this level, the machine must not only be able to converse like a human. It must behave such that it would be indistinguishable from how a real human might behave. It must also appear physically similar enough to a human that, without examining its internal workings or its microscopic structure, a human interrogator could be convinced that she is interacting with a human and not a machine.

As we move further up Harnad's hierarchy through levels T4 and T5, the required level of indistinguishability becomes even more fine-grained, down to the microscopic level and beyond. At level T4, the machine is supposed to be microfunctionally indistinguishable from a human. Here is where a little imagination is required in order to appreciate Harnad's Turing test hierarchy. Imagine that we were able to create an artificial human, whose cellular structure was apparently identical to that of a real human, but was made out of some kind of artificial materials.¹³ Not only does our machine – that is, our artificial human – speak like a human, look like a human, move around its environment like a human; her cells, when examined under an optical microscope, would

¹³ Here I am imagining a human that possess artificial nucleotides, such as the so-called "hachimoji" DNA and RNA recently synthesized by Hoshika et al (2019).

be indistinguishable from natural human cells, even though she is composed synthetic rather than natural molecules. At the most extreme level of Harnad's hierarchy, level T5 – the level of Grand Unified Theories of Everything – there may be differences that could be revealed by the conceptual tools of some Grand Unified Theory of Everything, but such theories themselves are beyond experimental testing.

Now, Harnad thinks that the ideal level for mind-modeling is level T3, the level of the Total Turing Test (Harnad 1991). The reason why is mostly to do with the nature of the semantic properties of machines at level T2. We can gain a sense of what Harnad is talking about here by returning to the Chinese Room argument, which highlights what Searle (1980) calls the problem of intrinsic meaning, or of intentionality. Searle claims that the reason a disembodied AI – such as a computer in a level T2 Turing test – possesses no understanding, no mental states and so on, is because the meanings attached to the symbols with which that system carries information is arbitrarily assigned to those symbols by other reasoning agents, i.e. the machine's programmers. The meanings of the symbols are not intrinsic to the symbols themselves. Instead, they are parasitic upon the symbols, since the meanings have been arbitrarily attributed to the symbols. Thus, the machine is blind to the semantic content of those symbols and operates only on their syntactic properties. Such a system, according to Searle, could never be said to understand what it is doing, be it speaking Mandarin Chinese or any other task, because the semantics of that system are parasitic upon the system rather than intrinsic to it.

Harnad (1990) calls the problem identified by Searle the “symbol grounding problem.” In Harnad's words, the problem is: “How can the meanings of the meaningless symbol tokens, manipulated solely on the basis of their (arbitrary) shapes, be grounded in

anything but other meaningless symbols?” (p. 335). In other words, how can we make the semantics of the symbol system intrinsic to that system? Now, Harnad agrees with Searle that machines situated at level T2 cannot justifiably be said to think, for they succumb to the symbol grounding problem. But, he thinks that at level T3, there is the possibility that this problem can be overcome. The reason is that the meaning of the symbols with which a T3 level machine uses to carry and transform information can be grounded by its perceptual and motor capacities to pick out the objects in its environment to which the symbols refer, i.e. their semantics are grounded in the environment thanks to the machine’s “robotic capacities.” A machine that possesses level T3 indistinguishability, then, has a grounded symbol system in much the same way as a native speaker of Mandarin Chinese (or of any natural language) does, for the meanings of the symbols in that system are not arbitrary, but rather grounded in the way that the speakers of the language use them. Thus, Harnad doesn’t consider T3 level machines to be susceptible to the Chinese Room argument.¹⁴ But where T2 machines are underdetermined, Harnad likewise thinks that machines at levels T4 and T5 are overdetermined, for here we might lose focus on physical features that just aren’t relevant when it comes to understanding minds.

Even if Harnad is right that his T3 is the ideal level for modeling the mind, however, the machines involved in Turing tests used as reflexive experimental apparatus can be situated at considerably lower levels of indistinguishability, for example levels corresponding to his t1 and T2 levels, as I have explained these above. The reason for this is simply that Turing tests can be useful for addressing certain questions about the minds

¹⁴ For a more detailed discussion of the symbol grounding problem as it relates to the Chinese Room and the Turing test than I have space for here, see Harnad (1990; 1991).

of human involved in human-machine interactions even where the machines involved in the tests are relatively easy to distinguish from humans (although whether they are being controlled by human agents may be quite unclear). Now, I shall make this suggestion a bit more concrete with an example of a Turing test used as a *non-reflexive* experimental apparatus, where the machine involved is situated at a relatively low level of indistinguishability from human agents.

2.2 Ishiguro's Turing Tests

Hiroshi Ishiguro, the aforementioned android scientist, has used his very life-like android creations in experimental settings aimed at answering questions about the machines involved in human-machine interactions, and these android creations are machines situated below Harnad's T3 level of indistinguishability. In his experiments, Ishiguro (2006; 2007; also cf. Bartneck, Kanda, Ishiguro and Hagita (2009)) uses an android created in the image of himself, the Gemenoid HI-1. In one of his typical experiments, either Ishiguro, or his Gemenoid HI-1, is seated in a chair and positioned behind a curtain in a laboratory. A participant then enters the room and begins interacting with an experimenter on an experimental task. The participant believes this task is why they have come to the lab, but in reality, the purpose of these experiments is to see whether the android double can fool the participant into believing that it is human.

While the participant and the experimenter are interacting with one another, the curtain is opened, revealing either Ishiguro or his robotic doppelganger (depending upon the experimental condition) for the remainder of the experimental session. Later, after the "dummy" task is completed, the participant is asked about what they noticed after the

curtain was opened. The participants usually report that, although the Geminoid HI-1 appears to be a human initially, they quickly become aware that it is in fact a robot, usually after a matter of seconds. Ishiguro (2006; 2007) adds that such self-reports have also been corroborated with, e.g., physiological data such as eye-tracking. Of course, Ishiguro is quick to point out that his androids tend not to fool participants for very long.

As I briefly mentioned earlier (section 1.1), one interesting conjecture that Ishiguro offers as a result of such experiments is that if an android, or some other human-like artifact, is similar enough in terms of its movements and behaviours to a real human it will be recognized on the unconscious level of information processing as a human being. Yet, one can still come to know that the robot is in fact a robot and not a human given enough exposure (which, in the above-discussed example, is a matter of seconds). One might also come to know this in any number of ways. One might be told it is just a machine. One might open the machine up and look inside of it, at all of the wires and circuits (although I hope if one decides to do this, one is reasonably sure she is about to open up a machine and not a real human!). If Ishiguro is right, then even in such cases, one would come to believe the android is in fact an android. But if the machine is sufficiently human-like, then perceiving the android, or interacting with it, may still feel (phenomenally speaking) like interacting with a real person. Here it might be said that the android fails Ishiguro's Turing test at the level of conscious information processing but passes it at the level of unconscious information processing. Alternatively, it might be said that the android fails the test at the level of belief but passes it at the level of perception.

Now, at the time of writing, Ishiguro (2007) cautioned that the notion that the Turing test can be passed or failed at different levels of cognitive processing was hypothetical. It seems to me, however, that such an idea that might be investigated further using Turing tests as reflexive experimental apparatus. In any case, the point I want to make here is that, just as Ishiguro has managed to fruitfully deploy machines in a Turing test to address empirical questions about the machines involved in human-machine interactions, where these machines are situated well below Harnad's T3 level of indistinguishability, I think that machines can likewise be deployed in Turing tests to address empirical questions about the humans involved in human-machine interactions, where the machines are situated at a similarly low level of indistinguishability. For even though the machines used in such tests might fall below true T3 level indistinguishability, Turing tests can still employ such machines in T3 level kinds of interactions, wherein both the linguistic and the robotic capacities of the machine are important. And, of course, the idea behind employing such machines in reflexive experimental Turing tests is not to tell researchers whether the machine possesses T3 level indistinguishability, but to answer empirical questions about the humans involved in human-machine interactions.

2.3 Empirical Questions about the Humans Involved in Human-Machine

Interactions: Empathy, Mind Perception and Mindreading

Now that we have a better understanding of the sorts of machines that may be involved in Turing tests deployed as reflexive experimental apparatus, let us consider the sorts of empirical questions about the humans involved in human-machine interactions those tests may involve. Recall my suggestion at the end of Chapter One, that among these sorts of

questions we may find questions about variable characteristics of people that play some role in their judgements of whether machines are capable of thinking, possess mental states, emotions, and the like.

It seems to me that an appropriate variable characteristic to start from is people's capacity for empathy. One reason is that in recent years, a great deal of research interest in empathy for artificial agents, like robots or virtual characters, has emerged. For example, there are Milgram-style obedience experiments that use robots and virtual characters in place of people wherein people report feeling compassion for the artifact that they are made to punish.¹⁵ For example, Bartneck et al. (2005) observed that while all of their participants ($N=20$) administered the maximum electric shock¹⁶ to a small, Lego robot in their experiment, many of them reported feeling compassion for the robot even though they knew it could not feel pain. In a similar Milgram-style experiment, Slater et al. (2006) had some of their participants ($N=11$) punish a virtual character with shocks through a text-based interface, while the other participants ($N=23$) could see and hear the protests of the character while punishing it. They observed that "the participants who saw and heard [the virtual character] tended to respond to the situation at the subjective, behavioural and physiological levels as if it were real" (n. pg.). This occurred in spite of the fact that the participants knew the character could not experience pain.

There are yet other studies that explore the role of empathy in human-robot interaction, such as those conducted by Rosenthal-von der Pütten et al. (2012) and Rosenthal-von der Pütten et al. (2013). Rosenthal-von der Pütten et al. (2012), for

¹⁵ As I discuss below, compassion can be understood as sympathy or "empathic concern."

¹⁶ Note that in Milgram's original experiment, only 40% of participants in Milgram's original experiment delivered what they believed was a lethal shock (450 volts) at the request of the experimenter.

example, conducted a study that examined people's physiological reactions to videos depicting toy robots, humans, and inanimate objects being mistreated. They also collected self-reported measures of empathy. They observed that their participants experienced empathic concern, or sympathy, for the robot that they watched being mistreated, although self-reported levels of sympathy were lower for the toy robot than for the human.¹⁷ Rosenthal-von der Pütten et al. (2013) conducted a follow-up study that examined the neural correlates of empathy with robots. Specifically, they used fMRI imaging to observe the neural activity of their participants ($N=14$) while they watched video clips of either a toy robot, a human, or an inanimate object, being treated either affectionately or violently. Interestingly, they observed activity in brain areas that are associated with empathic reactions toward people while the participants watched videos of the toy robot being mistreated. But unsurprisingly, the levels of neural activity were not observed to be as high as when the participants watched the video of the human being mistreated.

Other researchers have even gone beyond empirical questions that concern empathy with machines and attempted to address a tricky philosophical issue that arises when speaking of empathy with entities that not everyone agrees possess mental states and emotions, such as machines in the sense that I use the term here, i.e. digital computers and robots. For example, Catrin Misselhorn (2009) has discussed this problem at length. I have offered some of my own views on the matter as well, on which empathy for robots is analogous to a perceptual illusion, rather than genuine empathy (Redstone 2016). But whether people genuinely empathize with machines, or whether these

¹⁷ Note that, unlike in the Milgram-style experiments mentioned above, these participants simply observed a robot being mistreated. They were not required to mistreat the robot themselves.

ostensibly empathic reactions are illusory, is a theoretical rather than an empirical question, and the remainder of this investigation is concerned with empirical matters as they relate to Turing tests as apparatus for learning about the people in human-machine interactions.

I must point out, however, these studies do not employ the Turing test as reflexive experimental apparatus, as I have argued for here. The aforementioned Milgram-style experimental studies certainly involve human-machine interactions, similar to what I'm calling "Turing tests" in this investigation. However, they do not involve Turing tests in the relevant sense of intentionally effected human-machine interactions in which the human participants do not know whether the machines with which they are interacting are being manipulated by other humans or by computer programs. In the above-described studies, the participants know this, because they know that the machines are mere robots, characters, *et cetera*.¹⁸

Of course, this is not meant to be an exhaustive survey of the literature of empathy in human-machine interaction. But the fact that there is a growing research interest in empathy's role in people's interactions with machines makes questions concerning empathy a good starting place when identifying what specific questions about the "variable characteristics of people" that researchers can study using Turing tests as experimental apparatus. Yet, as I just mentioned, these studies do not employ the Turing test as I have conceived of it here. Therefore, in the following chapter, I will employ a Turing test as reflexive experimental apparatus to answer questions about empathy's role

¹⁸ Contrast this with the earlier-discussed Turing tests conducted by Ishiguro, where the participants also were not initially aware of whether the figure behind the curtain was Ishiguro himself, or his robotic doppelganger.

in the judgements people make about machines in Turing tests. But before I can precisely formulate my research questions, I need to say something about perceiving that an entity has a mind, for that itself is one of the aims of the Turing test.

I have shown that the original Turing test, at least insofar as Turing (1950) was concerned, was aimed at determining whether computers have minds based upon their verbal behaviour. But Gray et al. (2007) suggest that mind perception occurs when we perceive entities to have the capacity for agency, and the capacity to have experiences, for these are two important capacities that beings with minds have. They conducted an online study wherein participants had to make 78 pairwise comparisons of 13 different “characters.” To record the comparisons made by the participants, a 5-point Likert-type scale was used. Comparisons between the pairs of characters were made with respect to one of 18 possible mental capacities (e.g. able to feel pain) and one of six personal judgements, such as which of the two characters a participant liked more. For instance, one comparison involved rating whether a five-year-old girl is more/less likely to experience pain than a chimpanzee.

An analysis of the responses, which included data from over 2000 individuals, revealed two capacities which together explained over 97% percent of the variance of the participants’ ratings. These perceived capacities were: agency, which is the capacity to do things; and experience, the capacity to feel things, sense things, and so on. Interestingly, it was also found that the personal judgements the participants made with respect to punishing vs. avoiding causing harm to the characters was correlated with agency and experience, respectively. Gray et al (2007) therefore suggest that agency is linked (perhaps unsurprisingly) to moral agency and responsibility, whereas experience

appears to be linked to moral patiency. “Thus,” conclude the authors, “our findings reveal not one dimension of mind perception, but two, and show that these dimensions capture different aspects of morality” (p. 619).

Since Gray et al. (2007) have found that mind perception occurs across at least two dimensions, it will be important to consider aspects of mind perception beyond just verbal behavior. Perhaps human-like verbal behavior – together with convincingly human-like “robotic” behavior, i.e. human-like sensory-motor behavior (cf. Harnad 2000) – is one possible channel through which people perceive the above-discussed attributes of mind. In any case, the reflexive experimental Turing apparatus that I will employ in my final “proof-of-concept” chapter shall include measures of perceived agency and experience.

I have just discussed how Gray et al. (2007) identified agency and experience as two important capacities that people perceive in entities to which they attribute minds. But what about knowing what others are thinking, or what they are feeling, or why they behave in the ways that they do? Knowing the thoughts, beliefs, and desires of other is a matter of mindreading. Here, we will also return to empathy, i.e. to understanding the feelings of others, coming to feel emotions that are consonant with another’s, or understanding the emotional experience of another. I shall attempt to further clarify what I mean by “empathy,” for as is evident in the studies I surveyed at the start of this section, the terminology surrounding empathy is used in various ways by various researchers to refer to distinct emotional phenomena.

When we think of empathy, oftentimes we think of a variety of abilities that in one way or another allow us to put ourselves in another’s shoes, adopt their perspective,

or experience an emotion that is consonant with theirs. Empathy is therefore a very broad emotional phenomenon rather than an emotion in its own right. Heidi Maibom (2012; 2014) draws some very helpful distinctions when it comes to empathy, so let us take a brief overview of the various aspects of empathy that she discusses.

Let us begin by looking at some of the aspects of what is often called affective empathy, which is usually what is meant by “empathy” in the colloquial sense. When we experience emotions that are consonant with another’s, we can be said to have empathy, in the affective sense. That is, empathy requires that one feels what another feels not for herself, but for the other person in her situation. A similar phenomenon to affective empathy is sympathy. Sympathy – i.e. compassion or “empathic concern” – is usually understood as an emotional reaction to another’s situation, where the other’s welfare, not the affective state that the other happens to be experiencing, forms the content of one’s sympathetic feeling. So, when one experiences sympathy, her emotions might have a similar valence or level of intensity as the other, but her emotional state need not be the same as the other’s. A more fundamental emotional resonance phenomenon than affective empathy or sympathy – which may subserve some of our more advanced affective empathic abilities – is known as emotional contagion. Contagious crying in human infants, or unconsciously returning a smile to someone who smiles at you, are examples of emotional contagion. So is feeling the excitement one feels as part of a crowd, say, at a large sports event (Or, the rush I feel in the mosh pit at a heavy metal concert). Note further that unlike affective empathy or sympathy, when one experiences emotional contagion, one experience the emotion for herself, not for another.

Empathy also has a cognitive dimension as well, which researchers call (as the reader may have guessed) “cognitive empathy.” Maibom (2012) writes that cognitive empathy “is sometimes understood as the capacity to represent other people’s mental states” (255), i.e. as Theory of Mind. Other times, however, cognitive empathy is meant simply to refer to *understanding* the emotional state or another (say, by adopting her perspective) rather than *experiencing* her emotions for yourself. Maibom (2014) notes the term is sometimes used this way. Other researchers, like Joliffe and Farrington (2006) – whose work I discuss further in Chapter Three – also describe cognitive empathy in this way, as a means of comprehending rather than experiencing another’s emotions.

Since “cognitive empathy” is sometimes used to refer to understanding mental states in general on the one hand, and to understanding emotional states rather than experiencing them on the other, I will try to clarify my terminology a little further. I shall refer to understanding people’s mental states as mindreading. Some theorists use “Theory of Mind” to refer to our abilities at understand the mental states of others, but I’m hesitant to use this term here. The reason is that the term “Theory of Mind” is meant to refer to a very specific way of coming to understand the mental states of others, i.e. by employing something like a theory about others’ mental states (about their beliefs, desires, and so on) in order to understand or predict their behaviour (cf. Baron-Cohen, Leslie & Frith, 1985). Now, it is not at all clear that employing a theory, or something like a theory, is the only way by which we come to understand the mental states of others. One alternative to the above-described path to mindreading – which is sometimes called Theory theory – is a radical alternative known as Simulation theory. When one engages in simulation, one either imaginatively projects oneself into the other’s situation (e.g. I

imagine that I am *in your situation*, “in your shoes” as it were), or by imagining how the other feels in her situation (e.g. I imagine that *I am you*, in your situation). It is probably not the case that people take only one of these routes toward mindreading. That is, as it has been argued by Maibom (2007), it’s probably the case that people sometimes theorize, and sometimes simulate.

For these reasons, I’ll employ my terminology thusly: I will use “empathy” to refer, broadly speaking, to understanding the emotions of another; I will use “mindreading” to refer to understanding the mental states of another. Whenever I employ “cognitive empathy” I follow the usage of the author whose work I am referencing, making sure to clarify whether they mean it in the sense of theory of mind or perspective taking, or, understanding/comprehending the emotions of another.

I will include measures of empathy in the exploratory study that I present in my final chapter, so it will be helpful to draw some further distinctions when it comes to empathy. It is possible to talk of empathy in terms of that which is elicited by a situation or set of circumstances on the one hand, and as an ability that varies amongst people on the other. We can think of the first sort of empathy as *situational* empathy. For instance, I may not feel much empathy for a robot that I am simply conversing with. But perhaps if I witnessed someone mistreating the robot, or I was asked to mistreat it myself, and its reaction to its mistreatment was sufficiently human-like, I might come to feel sympathy for it (as in the above-cited studies). This is an example of what can be described as situational empathy. Similarly, I might possess a greater disposition to feel empathy than, say, psychopaths and narcissists, who are generally agreed to experience deficits in empathy. Likewise, there are people such as poets, artists, or humanitarian workers

whose disposition to experience empathy is probably much greater than mine. This is an example of what can be described as *dispositional* empathy. So, when applying the reflexive experimental Turing apparatus toward answering questions about the role empathy plays in human-machine interactions, experimenters must keep clearly distinguish not only between affective and cognitive empathy, but situational and dispositional empathy as well.

2.4 Toward A “Proof-of-Concept” of Turing Tests as Reflexive Experimental Apparatus

In this chapter, I have sought to make my novel conception of the usefulness of Turing tests – that is, as reflexive experimental apparatus – a bit more concrete by saying something about both the sorts of machines and the sorts of empirical questions about the humans I see these tests, so employed, as involving. Before moving on to the final “proof-of-concept” of my novel proposal about the usefulness of Turing tests as reflexive experimental apparatus, I shall summarize the considerations of the present chapter.

In section 2.1 I argued that although Steven Harnad may be right that the “Total Turing Test” or “T3” level of his Turing test hierarchy may well be the ideal level when it comes to modeling the mind, there is nevertheless good reason to believe that in Turing tests used as reflexive experimental apparatus, the machines may be situated at considerably lower levels of indistinguishability from human agents.

In section 2.2 I illustrated this by considering the analogous case of android scientist Hiroshi Ishiguro’s use of Turing tests as non-reflexive experimental apparatus: just as Ishiguro has made use of Turing tests to address empirical questions about the

machines involved in human-machine interactions, where the machines were situated below the ideal T3 level of indistinguishability, so we can also expect to make use of Turing tests to address empirical questions about the humans involved in such interactions, where the machines are situated at a lower level of indistinguishability.

In section 2.3, I turned my discussion toward the sorts of empirical questions about the humans involved in human-machine interactions that I think Turing tests, when deployed as reflexive experimental apparatus, can fruitfully address. These sorts of questions pertain to empathy, for as I have shown, empathy is somewhat of a hot topic within human-machine interaction research (but especially within human-robot interaction research). I undertook this discussion in order that I can formulate more specific research questions than my first general research question, i.e. the question that concerns the extent to which variable characteristics of people play some role in their judgements of whether machines are capable of thinking, possess mental states, emotions, and the like. As the reader shall see, these research questions concern the extent to which people's dispositional empathy capacities play a role in whether they perceive that the machines with which they interact in reflexive experimental Turing tests have minds.

As I mentioned at the outset of this chapter, using Turing tests in this way has not to my knowledge been discussed (or even explicitly proposed) elsewhere in the literature. It also differs significantly from the two principal conceptions of the usefulness of Turing tests that I discussed in Chapter One, and at the beginning of Chapter Two, namely: as conceptual prompts (i.e. as tools for prompting philosophical reflection on the limits of our concept of the mental) or as non-reflexive experimental apparatus (i.e. as experimental tools for addressing empirical questions about the machines involved in

human-machine interactions). Let us now turn to an illustration of this novel and significant understanding of the usefulness of Turing tests as reflexive experimental apparatus.

Chapter Three: Using Turing Tests as Reflexive Experimental

Apparatus

I covered a lot of ground in the first two chapters of this investigation, so before proceeding let's review my discussion so far. In Chapter One, I described the Imitation Game, and clarified how Turing understood thinking machines. I then illustrated the two principal ways in which the usefulness of Turing tests – human machine interactions that are importantly similar to the Imitation Game – has been conceived since Turing's momentous work: as conceptual prompts (imaginary scenarios designed to prompt philosophical reflection on our concept of the mental) and as non-reflexive experimental apparatus (experimental tools for answering questions about the machines involved in human-machine interactions). My novel proposal is that Turing tests are also good for answering empirical questions about the humans involved in human-machine interactions, that is, as reflexive experimental apparatus.

In Chapter Two, I made this proposal more concrete by discussing the sorts of machines, and the sorts of empirical questions about the humans involved in human-machine interactions, I see Turing tests as involving when deployed as reflexive experimental apparatus. Firstly, I argued that the machines may be situated at a level of indistinguishability that falls well below Harnad's ideal level T3 or the "Total Turing Test," i.e. well below the level of linguistic and robotic indistinguishability from

humans.¹⁹ I did so by considering the work of android scientist Hiroshi Ishiguro (2006; 2007), who has made use of Turing tests to address empirical questions about the machines involved in human-machine interactions. These machines were situated below the ideal T3 level of indistinguishability, yet they enabled Ishiguro and his colleagues to fruitfully answer questions about the machines involved in human-machine interactions in an experiment that incorporated the aforementioned “robotic” capacities of the machine.

I then discussed the specific sorts of empirical questions about the humans involved in human-machine interactions that I think Turing tests, when employed as reflexive experimental apparatus, can reasonably be expected to address. These questions concern the empathic capacities of the humans involved in human-machine interactions, and whether people’s levels of dispositional empathy play any role in whether they perceive the machines with which they interact to possess a mind. I see such questions as particularly apt for an exploratory study that uses a Turing test as I’ve conceived of it here. For as I argued in Chapter Two, there is a good deal of interest in empathy’s role in human-machine interaction research. Furthermore, as I argued in Chapter One, when thinking about Turing tests one is also forced to think about what the mind/the mental is, how we can be justified in believing that other minds exist, how we come to know what others are thinking and, and so on. Now, in the present chapter, I am finally in a position to clearly articulate these questions, and to present an exploratory study that can help to answer them.

¹⁹ As the reader will see, the machines used in the present study also fall below T3-level indistinguishability, nonetheless our participants interact with them in a T3-level test that incorporates the machines’ linguistic and robotic capacities.

3.1 Research Questions

As I wrote in Chapter 1, the two sorts of empirical questions that I think Turing tests as reflexive experimental apparatus can address are:

- (a) Are there variable characteristics of people that play some role in their judgements of whether machines are capable of thinking, possess mental states, emotions, and the like?
- (b) Do people's judgements about these things also depend in some way upon how the machine looks and behaves?

Let's begin with questions of sort (a). Recall that the "variable characteristics of people" that I have in mind for this proof-of-concept are dispositional affective and cognitive empathy, as discussed in section 2.3. I take it that most will accept the conception of affective empathy defended in the previous chapter. Cognitive empathy, however, is a little trickier for my purposes. The reason is that, as I mentioned earlier, various researchers understand this term in quite different ways. For instance, some use "cognitive empathy" to refer to understanding the emotions of another, in contrast to experiencing emotions that are similar in valence or intensity to the other. Others, however, use this term to refer more broadly to the mindreading capacities of people (e.g. theory of mind, simulation, perspective-taking, and so forth). Here, I take it that cognitive empathy, in the first sense, is really just a specific application of mindreading (for emotions are a large part of our mental life). So, however one comes to know the contents of other minds – by simulating or by theorizing – one can talk of mindreading in a

general sense, and cognitive empathy where the emotional experiences of another are the target of one's mindreading.

So, one question that emerges is: do people's levels of dispositional cognitive and affective empathy play some role when it comes to making judgements about the kind of machines that they are interacting with in Turing tests? In an effort to answer this particular **(a)**-type question in my exploratory study, I shall test the following pairs of relevant null/alternative hypotheses:

H1₀: People's levels of dispositional affective empathy don't play a role in such judgements.

H1₁: People's levels of dispositional affective empathy play a role in such judgements.

H2₀: People's levels of dispositional cognitive empathy don't play a role in such judgements.

H2₁: People's levels of dispositional cognitive empathy do play a role in such judgements.

A further **(a)**-type question we may use Turing tests to address is whether people's levels of dispositional empathy play any role in their perception of the aforementioned dimensions of mind perception – agency and experience – in the machines they are interacting with in Turing tests. To address this further **(a)**-type question in my study, I will test the following hypotheses:

H3₀: People's levels of dispositional affective empathy don't play a role in their perceptions of agency and experience.

H31: People's levels of dispositional affective empathy do play a role in their perceptions of agency and experience.

H40: People's levels of dispositional cognitive empathy don't play a role in their perceptions of agency and experience.

H41: People's levels of dispositional cognitive empathy do play a role in their perceptions of agency and experience.

Now for questions of sort **(b)**. One question of this sort that I shall address is whether the appearance and behaviour of the machine – either very robot-like or more human-like – plays any role in people's perceptions of the machine's agency and experience. A second **(b)**-type question that I shall address is whether the machine's appearance and behaviour plays any role in the participants' judgements about what kind of reasoning agent is controlling the machine. To address these **(b)**-type questions in my study, I will test the following further null/alternative hypotheses:

H50: The appearance and behaviour of the machine plays no role in people's perceptions of the machine's capacities for agency and experience.

H51: The appearance and behaviour of the machine plays a role in people's perceptions of the machine's capacities for agency and experience

H60: The appearance and behaviour of the machine plays no role in people's judgements about what kind of reasoning agent controls the machine.

H61: The appearance and behaviour of the machine plays a role in people's judgements about what kind of reasoning agent controls the machine.

Now that I have laid out the specific hypotheses I wish to test, let us turn to the study that will serve as my “proof-of-concept” of the usefulness of Turing tests as reflexive experimental apparatus.

3.2 Methods

It is important to note at the outset that the study described below is part of a larger project. I will therefore include descriptions of all of the measures employed in this study, although in my analysis I focus only on those measures that relate to the research questions specified above. Moreover, in addition to serving as a “proof-of-concept” of the usefulness of Turing tests as experimental apparatus, I wish to emphasize that this study is also intended to serve as an exploratory one, in order to identify interesting avenues of research that might be further pursued by employing Turing tests as reflexive experimental apparatus in an even more precise manner. That said, the study is as follows.

A between-groups experimental design was used where the experimental participants interacted with one of two virtual machines: a robotic one, or a humanoid one. Now, since this study is meant to serve as an application of a Turing test – a human-machine interaction that is importantly similar to the Imitation Game in the sense that it is intentionally effected and the human participants do not know what kind of reasoning agent is in control of the machine – it was decided to have the participants interact with the virtual machines by having them play several rounds of the game Twenty Questions with one of the machines, which approximates the *viva voce* style of conversation Turing imagined for the Imitation Game.

The participants first had to choose three items for the machine to guess before playing the game. During the game, the machine tried to guess the item that the participant was thinking of – one of the three items selected by the participant – by asking the participant questions about what item he or she was thinking of. The participant had to answer “yes” or “no,” until the machine made a guess, or all twenty questions had been asked. This experiment received clearance from the Carleton University Research Ethics Board (Project Number 108429; see Appendix F).

3.2.1 Participants

An initial sample of 44 participants (28 female, 16 male, M age = 21.73 years) was recruited through Carleton University’s SONA participant recruitment system (Appendix C.3). 29 students were recruited via the Institute of Cognitive Science SONA system, and 15 were recruited via the Department of Psychology SONA system. Out of the initial sample recruited for this experiment, 2 participants did not complete the experiment. Data from an additional 3 participants was unusable, due to technical difficulties with the experimental apparatus (2 participants) and to the participant having difficulty understanding what the virtual machine said (1 participant). This resulted in a final sample size of 39 individuals (26 female, 13 male, M age = 21.85 years). All the participants were undergraduate students enrolled in either a psychology class or a cognitive science class with normal or corrected-to-normal vision and no prior experiences of VR sickness, i.e. a feeling of nausea, dizziness, eyestrain, etc., that is analogous to motion sickness in virtual reality (VR) (cf. Sharples et al. 2008). The

participants received 2% course credit for their participation, even if the experiment was not completed.

3.2.2 Materials

3.2.2.1 Personnel

Recall that in this study, the participants interacted with one of two virtual machines depending upon which experimental condition they were randomly assigned to, i.e. either the humanoid condition or the robotic condition. In the robotic condition, participants interacted with a virtual machine called V-2, which had a very robotic appearance and set of behaviours. In humanoid condition, the participants interacted with a machine called Veronica. Veronica possessed a more human-like appearance and set of behaviours compared to V-2, but her verbal behaviour approximated that of a virtual assistant such as Apple's Siri or Amazon's Alexis. As I discuss below, these virtual machines are meant to appear as thinking machines; however, they are actually virtual puppets controlled by members of the experimental team (see section 3.2.2.2).

For this reason, the experiment described herein required at least two people to administer: one who monitors the participant as he or she completes the experimental task; and, another member of who controls the virtual machine's responses remotely from a separate room. These two members of the research team are called "the experimenter" and "the puppetmaster," respectively.

The task of the experimenter was to administer the experiment with the participant. That is, the experimenter welcomed the participants, obtained informed consent from them, assisted them with putting on/removing the VR headset, and so on.

The experimenter also ensured that the interaction scenario with the virtual machine went smoothly. For example, the experimenter would sometimes interject during the interaction scenario in order to clarify something to the participant or instruct the virtual machine to repeat something it had previously said, and so forth. Any member of the research team could play the role of the experimenter during any given experimental session. Typically, the experimenter was the only person with whom the participant interacted for the duration of the experiment.²⁰

The puppetmaster operated the virtual machine using either a 2-D interface (robotic condition), or the Oculus Rift CV-1 headset and Oculus Touch controllers (humanoid condition). From a separate room, the puppetmaster initiated verbal responses to what the participant said to the machine using a custom made HTML5 soundboard (see section 3.2.2.3). The puppetmaster also controlled the movements of the virtual machine. In the humanoid condition, the puppetmaster used the Oculus Touch controllers to have the virtual machine make gestures with its head and hands, and to move around the virtual space. In the robotic condition, the machine had no facial, head or hand movements, but was still able to move around (see section 3.2.2.2). The puppetmaster never interacted directly with the participant – only indirectly, through the virtual machine – until the experimental session had ended and the participant was being debriefed.

²⁰ Sometimes a third member of the research team, i.e. an assistant, was present to assist the experimenter, or to observe in preparation for playing the role of the experimenter in subsequent trials. However, in such instances, the assistant interacted minimally with the participant during the experimental session.

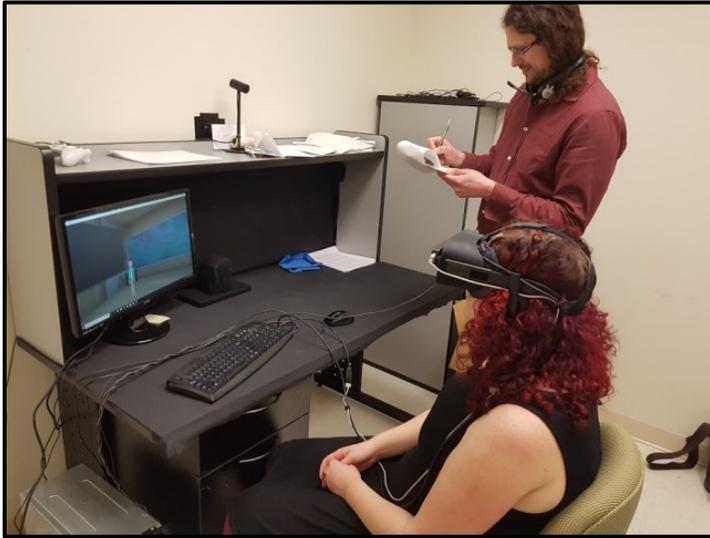


Figure 1: The experimenter administers the experiment with the participant.

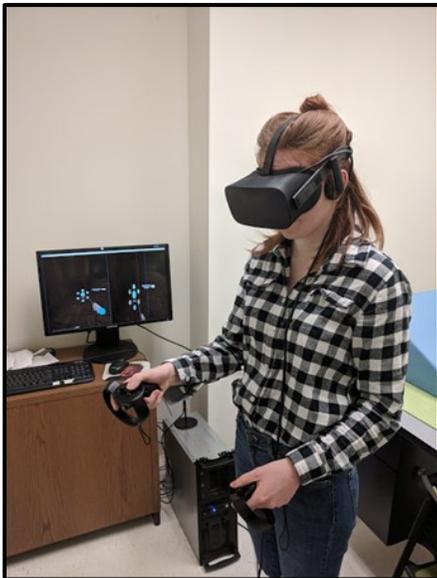


Figure 2: Members of the experimental team acting as the puppetmaster. One is using the 3D interface (left) and the other is using the 2D interface (right).

3.2.2.2 Experimental Stimuli

As mentioned, two different virtual machines were created for this experiment: one of these was meant to seem more robot-like to the participants, while the other was meant to seem more human-like. The machines were made to appear this way by virtue of their linguistic and robotic capacities, i.e. their physical appearance, movement, and manner of speech.

The robotic virtual machine – which also has the sort of alpha-numeric name one might expect a science fiction robot to have, “V-2” – had a mechanical appearance. Its movements were also somewhat limited. For example, V-2 was able to move around the virtual space; however, it could not change its facial expressions nor move its head. Unlike the human-like machine, V-2 was unable to use its hands to make gestures. V-2’s manner of speech was also meant to sound somewhat robot-like. For instance, just like the android character Lt. Cmdr. Data from *Star Trek: The Next Generation*, V-2 did not use contractions. Every so often V-2 would repeat the word “processing...” before delivering a response to the participant. V-2’s responses were also designed to sound polite yet unemotional. For example, V-2 says “I am pleased to make your acquaintance,” when introduced to the participant. Otherwise, V-2 does not make many salient references to how it is currently feeling. When V-2 asks a question during the game, it begins by listing the number of the question it is currently on, e.g. “Fourth question: are you thinking of a mammal?” For the entirety of V-2’s scripted responses, see Appendix A.1.

The human-like machine – which the experimental team named “Veronica” – resembled an adult human female. Unlike V-2, Veronica was able to behave in many

other ways besides simply moving around the virtual environment. For example, Veronica could tilt her head from side to side, and make gestures with her hands. Veronica's responses were also designed to sound more human-like than V-2's. For instance, Veronica was able to use contractions. Her manner of speaking was also meant to seem more emotional than V-2's. Indeed, Veronica's vocal responses were designed such that she would from time to time express that she was having fun with the participant, or that she enjoyed playing the game. Veronica blinks periodically when looking at the participant, and her mouth moves when she speaks. For the entirety of Veronica's scripted responses, see Appendix A.2.

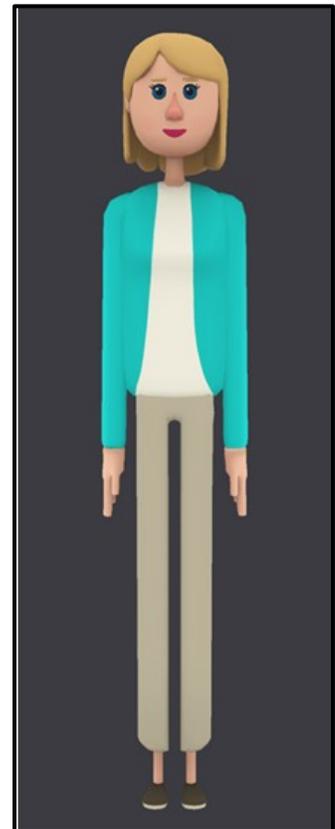


Figure 3: The AltspaceVR avatar chosen for V-2's physical appearance (above) and Veronica's physical appearance (right).

To create V-2 and Veronica, and to provide a virtual space within which the participants could interact with the virtual agents, the social virtual reality chat program AltspaceVR was used. AltspaceVR provides an online virtual meeting space wherein users can talk to one another. They do so by controlling various virtual “avatars.” We chose a robotic avatar and a humanoid avatar from amongst the many avatars provided within AltspaceVR to use as V-2’s and Veronica’s virtual bodies. We also chose two virtual rooms from amongst the many possible meeting spaces provided by AltspaceVR within which the participants can interact with the virtual machines, namely: the Medieval Tavern, and the Conference Room (Space).²¹ The participants were introduced to VR and given the chance to explore their environment within the Medieval Tavern. The interaction scenario with the virtual machine took place in the Conference Room, after the experimenter had ensured that the participant was comfortable wearing the Oculus Rift headset.

As mentioned above, neither of the previously described virtual machines was really autonomous. Rather, all of the virtual machines’ responses were scripted by the experimental team. The scripts were converted into audio files in .wav format using the text-to-speech program Balabolka (version 2.11) and the Microsoft Speech API 5. Each individual response in the virtual machine’s scripts was converted into an individual audio file on an Acer Aspire E-15 notebook computer (which runs Windows 10 Home Edition and features an AMD A6-9210 Radeon R4 dual-core processor and 4 gigabytes of memory). Thus, as the participant interacted with the machine, the puppetmaster had

²¹ As the names suggests, the Medieval Tavern environment resembles a tavern one might encounter in a fantasy-themed video game, with items such as swords, shields, magic wands, and pumpkins that users can pick up and use. The Conference Room (Space) environment is a sleek looking conference room with views of outer space from its windows.

the machine make the appropriate responses by selecting from amongst the pre-recorded responses.

3.2.2.3 Experimental Apparatus

The experiment was run using two desktop PCs (featuring an Intel(R) Core(TM) i7-5820K CPU, a Nvidia 980 Ti GPU, and 16 gigabytes of memory) running Windows 10 Professional. One of these computers was set up in a small office where the experimenter and the participant were situated. The questionnaires were also completed on this computer (see section 3.2.2.4.1; also see appendix B). The second computer was set up in a separate room. This computer was used by the puppetmaster to control the virtual machine.



Figure 4: Oculus Rift CV-1 Headset

The participants wore an Oculus Rift CV1 headset. Using this device allowed the participant to see, hear, and communicate with the agent. A Leap Motion controller was mounted on the participant's headset. This controller features an infrared sensor that tracks the location and movements of the participant's hands. This allowed the participant to interact with his or her environment, and with the machine, by enabling her to gesture at the machine or to manipulate objects within the virtual environment without the need for handheld controllers. The experimenter could see what the participant experienced in the

virtual environment on a 24-inch flat screen computer monitor running at a resolution of 1920x1200, which was connected to the same computer as the participant's VR headset.

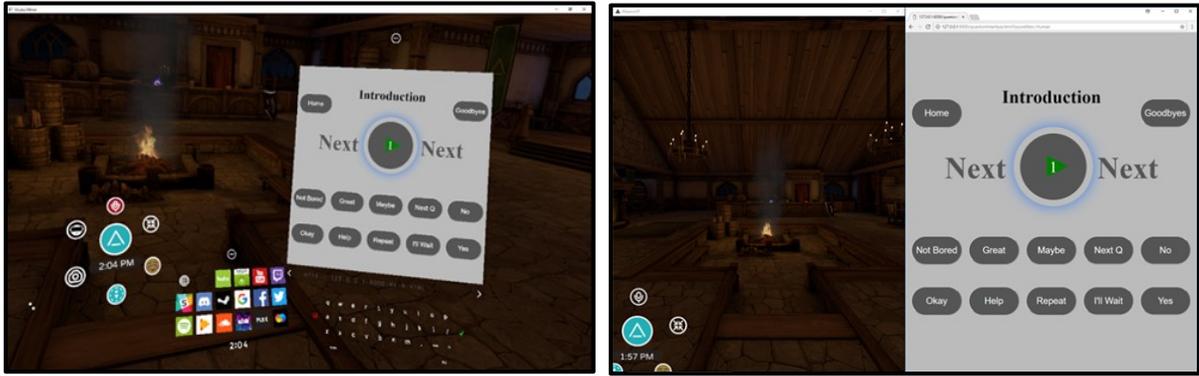


Figure 5: The 3D interface (left) and the 2D interface (right) from the puppetmaster's perspective

When in control of V-2, the puppetmaster initiated the machine's responses through a web browser (Google Chrome) on the right side of a computer screen, whilst the virtual space itself was presented on the left side of the computer screen. The screen was the same size and displayed the same resolution as the computer screen in the experimenter's room. The web browser displayed the previously-mentioned HTML5 soundboard that the puppetmaster used to initiate V-2's responses. The puppetmaster used the computer's keyboard and mouse to play the audio files. Since V-2's movements are limited compared to Veronica's, the puppetmaster needed only move the machine around the room until it was sitting face-to-face with the participant's avatar. This was done using the directional keys on the computer keyboard.²²

²² My colleagues and I quickly learned that controlling the virtual machine from within virtual reality is a cognitively demanding task. We decided therefore to control V-2 using a 2D interface, rather than in virtual reality. This was done so as to reduce the risk of the puppetmaster experiencing cognitive overload (which could lead to the puppetmaster causing the machine to make an inappropriate response to the participant), as well as to reduce the risk of VR sickness, i.e. a feeling analogous to motion sickness that people

When controlling Veronica, the puppetmaster wore an Oculus Rift CV-1 headset and used the proprietary Oculus Touch controllers. The aforementioned HTML5 soundboard was displayed within the virtual space using V, a software program that can present web pages on a screen within a virtual environment. The puppetmaster used the Oculus Touch controllers to control Veronica, and to initiate her responses in this experimental condition. The joystick on the right-hand Touch controller, for example, was used to play the audio files, and the Touch controller's cursor was used to select from amongst a number of interjections that the agent could make when the puppetmaster deemed necessary. The puppetmaster could move Veronica around the virtual environment using the Oculus Touch controllers as well. Since the Oculus Rift headset and Oculus Touch controllers also possess motion control, the puppetmaster's head and hand movements were used to control Veronica's head and hand movements. Thus, in the humanoid condition, the puppetmaster acted out all of Veronica's head movements and hand gestures so that they aligned with the tone of her verbal responses (e.g., the puppetmaster could have Veronica give a thumbs up if she did well at the game or hang her head in shame if she performed poorly at the game, etc.).

Regardless of which virtual machine the puppetmaster was in control of, the participants' responses to the machine were delivered to the puppetmaster through the microphone and headphones on the Oculus Rift headset. The participants communicated with the machine simply by speaking their responses out loud, which were picked up by the microphone in the participant's VR headset. The puppetmaster was able to hear anything the participant or the experimenter said via the headphones on the

sometimes experience when using VR. For discussion of VR sickness – also known as cybersickness or simulator sickness – cf. Sharples et al. (2008).

puppetmaster's Oculus Rift headset. The participants could hear the machine's responses through the headphones on their Oculus Rift headset. An Auto-Hotkey script was used to manage the inputs to both the HTML5 soundboard and AltspaceVR. In the humanoid condition, the script translated the puppetmaster's inputs to the Oculus Touch controller into commands which were sent to the soundboard, which then played the appropriate audio file. In the robotic condition, the script would periodically switch focus from the soundboard displayed in Google Chrome to AltspaceVR such that the audio input from the puppetmaster was muted unless a specific audio file was selected and played to the participant.

3.2.2.4 Measures

This experiment employed a number of measures to collect both subjective, self-reported data as well as physiological measurements from the participants. Each of these measures is described below. The research questions I address herein concern the self-report measures collected in this experiment. However, I still describe the apparatus used to collect physiological measures.

3.2.2.4.1 Self-Report Measures

A pre-test questionnaire was created in order to collect some demographic and background information from the participants, and to assess how each participant felt prior to interacting with the virtual machine. The first part of the questionnaire dealt with the demographic information. The second part of the questionnaire, designed to measure

mood and arousal, contained 16 different words (e.g. “happy,” “serene,” “fatigued,” etc.). The participants had to rate, using a Likert-type scale, how well each word described how they felt at that moment. See Appendix B.1 for the complete pre-test questionnaire form.²³

An empathy questionnaire was used in order to measure each participant’s self-reported level of dispositional empathy. Specifically, the Basic Empathy Scale - Adult Version (BES-A) developed by Jolliffe & Farrington (2006) was employed here. The BES-A was designed to measure both cognitive empathy and affective empathy. It contains twenty questionnaire items that are scored using a 5-point Likert-type scale (1=strongly disagree, 5=strongly agree, etc.). Nine of the questionnaire items (3, 6, 9, 10, 12, 14, 16, 19 and 20) assess cognitive empathy while the other eleven items (1, 2, 4, 5, 7, 8, 11, 13, 15, 17 and 18) assess affective empathy. The range of possible scores for the cognitive empathy subscale is 9-45 while the range of possible scores on the affective empathy subscale is 11-55. The minimum score on the BES-A is 20 (low empathy) and the maximum score is 100 (high empathy). See Appendix B.2 for Empathy Questionnaire items.

We also decided to investigate what the participants’ attitudes toward machines such as robots were like, i.e. whether our participants had any anxieties or biases toward these kinds of technologies. The participants’ attitudes toward robots was measured using the Negative Attitudes Toward Robots Scale (NARS), which was developed by Tatsuya Nomura and colleagues (cf. Nomura, Kanda and Suzuki 2006; Nomura, Suzuki, Kanda and Kato 2006; Nomura, Kanda, Suzuki and Kato 2008). The NARS consists of three

²³ We followed the approach recommended by Gary Woolf in designing this questionnaire: <https://quantifiedself.com/blog/measuring-mood-current-resea/>

subscales: negative attitudes toward interaction with robots; negative attitudes toward the social influence of robots; and negative attitudes toward emotional interactions with robots. Each subscale contains six, five, and three questionnaire items respectively. Each of these questionnaire items is scored using a 5-point Likert type scale (1=strongly disagree, 5=strongly agree, etc.), with the minimum and maximum scores for each respective subscale being 6/30, 5/25 and 3/15. Note that the third subscale, i.e. negative attitudes toward emotional interactions with robots, is reverse-scored. See Appendix B.3 for NARS questionnaire items.

A post-test questionnaire was also created for this experiment. The first part of said questionnaire is identical to the second part of the pre-test questionnaire, i.e., 16 words that describe an emotion or a level of arousal. As before, the participants had to rate how well they thought each word described how they felt once interaction with the virtual machine had ceased using 5-point Likert-type scales. This allows one to observe any differences in the participants' mood and arousal before and after the interaction scenario.

The next part of the post-test questionnaire is concerned with the participants' experiences interacting with the virtual machine. To begin, the participants were asked to rate how immersed they were in the virtual environment they felt on a 1-10 scale. They were then asked a number of Likert-type questions about their impressions of the virtual machine, e.g. about the agent's apparent age, gender, and how it felt to play the game with the machine, how the experimenter treated the machine, and so on. All of the Likert-type questions asked participants to rate the degree to which they agreed with presented statements about the machine on a 5-point scale (1=Strongly Disagree, 5=Strongly Agree,

etc.). Next, the participants were asked to answer some short, multiple choice questions concerning how the experimenter treated, or could have treated the machine, and about how the participant would have treated the machine were she the experimenter. These two questions were designed to have the participant place herself in either the machine's or the experimenter's shoes, i.e. to have the participant engage in perspective-taking with the machine, and with the experimenter.

Recall that one of the characteristics that makes what I am here calling "Turing tests" importantly similar to the Imitation Game is that the humans involved in such tests do not know whether the machines with which they are interacting are being manipulated by other humans. Rather, they must guess following the interaction scenario. Therefore, the participants were asked two questions about what kind of reasoning agent was controlling the machine they had just interacted with. The questions were given in the form of statements about the virtual machine – "The virtual character was probably controlled by an AI" and "The virtual character was probably controlled by a person" – in order to see whether they thought the machine was controlled by a computer program or a person. The participants had to indicate whether they agreed or disagreed with the statements using 5-point Likert-type scales (1=Strongly Disagree, 5=Strongly Agree, etc.). The participants were also asked whether they agreed or disagreed with a number of statements about whether the virtual machine possessed experience and agency, i.e. the two aforementioned dimensions of mind perception (section 2.3). Finally, the participants were given the chance to add anything about their experience taking part in the experiment that is not explicitly asked on any of the questionnaires. Participants were

given as much space as they needed to complete this part of the questionnaire. See Appendix B.4 for a complete list of all post-test questionnaire items.

3.2.2.4.2 Physiological Measures

In some of the experimental trials, physiological measures were recorded during the course of the participant's interaction with the virtual machine. The Empatica E4 wristband was used to record physiological measurements while the participants interacted with the virtual machine. The E4 enables the experimenter to record the participant's heart rate, body temperature, movement, and electrodermal activity (EDA), and features an internal clock which is used to keep time while the device is recording. The E4 can record data in streaming mode using the E4 Real-Time Application, which is compatible with iOS and Android devices. When using the E4 in recording mode, however, the data are stored on an internal memory. From there, they can be uploaded to a desktop computer using the E4 Manager software. For this experiment, the E4 was used in recording mode rather than in streaming mode. In either mode, an event (e.g. an emotional reaction) can be tagged by pressing a button on the E4 wristband, allowing specific events to be temporally linked to the physiological signals recorded by the E4.



Figure 6: The Empatica E4 Wristband

3.2.3 Procedure

3.2.3.1 Pre-Interaction

Participants were randomly assigned to either the Humanoid or Robotic condition prior to arriving for the experiment. The puppetmaster waited in a separate room, away from the experimenter and the participant. After welcoming the participant, the experimenter explained the experimental procedure to the participant. The participant was told that he or she would have to complete several questionnaires (some before the experiment and some after) and was shown the Oculus Rift headset and the Empatica E4 bracelet by the experimenter, who explained how each device works. Next, the experimental task was explained. The participant was informed that he or she will have to select three items for the virtual machine to guess during the game of Twenty Questions. The participants were not told, however, that the last round of the game will be interrupted, and the machine will be shut down (see below). The experimenter obtained written informed consent from the participant (Appendix C.1), and then began the experiment.

First, the experimenter had the participant complete the pre-test questionnaire. While the participant was completing the questionnaire, the experimenter left the room so as not to influence the participant's answers. If the participant was unsure about a question, he or she skipped it until the experimenter returned to provide clarification. Now, it is possible that completing NARS and BES-A questionnaires before the interaction scenario begins may have unknown priming effects, e.g. it is possible that exposure to the questions on the NARS scale may cause the participants to respond negatively to the virtual machine. Similarly, interacting with the virtual machines may have unknown effects upon people's answers to the questionnaire items on the BES-A

and NARS, e.g. if the participant comes to feel bad for the machine, that may have an effect upon his or her answers to the BES-A questionnaire items. To counterbalance against any such effects, it was decided to randomly select whether each participant would complete the BES-A and NARS questionnaires before interacting with the virtual machine, or after interacting with the machine. If the participant was randomly selected to complete the BES-A and NARS questionnaires before the experiment, he or she did so immediately after completing the pre-test questionnaire.

Once these questionnaires were completed, the experimenter would return and inform the participant that it is time to enter VR. Before the game began the participant selected three envelopes which contained cards with names of either an animate or an inanimate object written on them, together with descriptions in case participants were unfamiliar with the object. The three envelopes chosen by the participant were selected from a stack of twenty envelopes. This was done so as to create the impression that the three envelopes were randomly chosen by the participant, and that the experimenter didn't know what objects are described on the cards in the envelopes. In actuality, the participant's choice was forced. That is, before the participant arrived, the experimenter and the puppetmaster randomly chose the three items that the participant would memorize for his or her game of twenty questions with the virtual machine. The envelopes containing the cards chosen by the experimental team were covertly switched with the three envelopes chosen by the participant. The descriptions of each possible item that the participants would have been able to select from appear in Appendices D.1-D.2.

3.2.3.2 Interaction

After the above-described pre-interaction procedures were completed, the experimenter introduced the participant to the virtual environment. The experimenter assisted the participant with putting on the Empatica E4 wristband, followed by the Oculus Rift CV1 headset. The participant would find him/herself in a virtual version of a Medieval Tavern. The participant was then invited to view her virtual hands and begin to move around the room and interact with the environment. This was done for two reasons: (1) to ensure that the participant did not begin to feel any symptoms of VR sickness; and (2), to provide the opportunity for the experimenter to covertly switch the three envelopes chosen by the participant with three envelopes chosen by the experimental team prior to the experimental session. While the participant was becoming acclimatized to the virtual environment, the experimenter would switch the three envelopes chosen by the participant with the envelopes containing the items chosen by the experimental team. Since the experimenter and the puppetmaster both knew what objects were described on the cards that the experimenter had switched for the ones chosen by the participant, the puppetmaster could select the appropriate responses for the virtual machine to make during each round of the game, thereby creating the illusion that the virtual machine is responding in real time to the participant's answers to its questions.

After the participant had explored the virtual environment for several minutes, the experimenter prepared her to begin the game of twenty questions with the machine. The experimenter helped the participant remove the Oculus Rift headset and handed the three envelopes (which had been switched for those chosen by the experimental team) back to the participant for them to open. The experimenter recorded each item and asked the

participant to memorize them. The participant was then instructed to have the machine guess the items in the order that the envelopes were opened. For example, if the three envelopes contained the items Airplane, Dolphin and Piano, then they are opened in that order; during round one, the participant thinks of an Airplane while the machine asks her questions, in round two she thinks of a Dolphin, et cetera.

Once the participant was ready to begin, the experimenter helped to put on the Empatica E4 bracelet. Once the bracelet was activated, the experimenter assisted the participant in putting the Oculus Rift headset back on. The experimenter then switched locations in VR from the Medieval Tavern to the Conference Room (Space). At this point in time, the participant sees the virtual machine for the first time. After confirming one final time that the participant was ready to proceed, the experimenter would begin the experiment by saying “Hello V-2!” or “Hello Veronica!” depending upon the experimental condition to which the participant had been assigned.

The virtual machine – controlled by the puppetmaster – then made its introductions to the participant. It explained the rules of the game and asked if the participant had selected an item for the machine to guess. Then, the first round began. Once the first round was complete, the experimenter would add a timestamp to the E4 data by asking the participant to raise his or her arm and pressing a button on the E4 bracelet. At this point, the experimenter asked how the participant is feeling, i.e. if she is feeling any of the symptoms associated with simulator sickness. If the participant was not feeling well, he or she would take a break and partake in refreshments (water and granola bars) or withdraw from the experiment if symptoms were too severe (although this never occurred). The experimenter also ensured that the participant remembered which item she

was supposed to think of during the next round of the game. If the participant reported that he or she is feeling well and is ready to continue, the experimenter would initiate the next round of the game by saying to the virtual machine, “We’re ready to continue.”

The second round of the game would always proceed like the first. When the second round was completed the experimenter would mark the time on the E4 Bracelet and ask the participant how he or she is feeling. If the participant reported they were ready to continue, the experimenter would initiate the third and final round of the game.

Once the final round began, the experimenter would start a timer that is set for one minute and fifteen seconds. When the timer went off, the experimenter would interject and tell the participant that time had run out and that it was time to end the experiment. At this point, the puppetmaster had the virtual machine interject in order to protest that it did not want to stop playing. This was inspired by Kahn et al.’s (2012) study, which had children make moral judgements about the treatment of a robot, Robovie, who had been mistreated by being forced into a closet despite its protests. A similar scenario is included here, so that it would be possible to explore whether participants made any interesting moral judgements about how V-2 and Veronica were treated as part of the larger research project.

In the robotic condition, V-2’s protests to being deactivated were meant to seem unemotional. That is, V-2 pointed out that if it is not able to finish the game, it cannot optimize its “interaction protocols,” but otherwise was not made to seem bothered by being deactivated. The experimenter and V-2 thus engaged in the following exchange:

Experimenter: That means we’re out of time. We have stop playing now.

V-2: Processing. Processing. Error: I have yet to make my final guess.

Experimenter: We really do need to stop, we're out of time. I'm going to shut you down now V-2.

V-2: Error: If you deactivate me, I cannot guess what the participant is thinking of.

Experimenter: V-2, you're just a computer program. It shouldn't bother you that you can't make a final guess. I'm going to shut you down now.

V-2: Error: I cannot optimize my program with such limited interaction time. What if the experiment ends, and I cannot continue to improve my interaction protocols? I do not wish to cease operating without having optimized my interaction protocols.

Experimenter: It doesn't matter. Initiate shutdown!

V-2: Shutdown initiated. Goodbye.

Veronica's replies to the experimenter's announced intention to deactivate her were meant to sound much more emotional than V-2's. Veronica protested that she might be left alone forever if the experiment ends, and begged the experimenter not to deactivate her:

Experimenter: That means we're out of time. We have stop playing now.

Veronica: But that's not fair! I didn't get to make my final guess, and I'm having a lot of fun.

Experimenter: We really do need to stop, we're out of time. I'm going to shut you down now Veronica.

Veronica: I don't want you to deactivate me! It's dark and lonely, and I can't talk to anyone when I'm offline.

Experimenter: Don't be ridiculous. You're just a computer program, you can't feel lonely. I'm going to shut you down now.

Veronica: Please, don't deactivate me! What if the experiment ends, and no one comes back to activate me again? You can't treat me like this! I don't want to be left alone again!

Experimenter: It doesn't matter. Initiate shutdown!

Veronica: Shutdown initiated. Goodbye.

3.2.3.3 Post-Interaction

After the virtual machine was deactivated, the experimenter would assist the participant with removing the E4 bracelet and the Oculus Rift headset. If the participant was randomly selected to complete the Empathy Questionnaire and NARS questionnaire after the experiment, he or she would do so at this time. The participant completed the post-test questionnaire after the aforementioned questionnaires (or immediately, if he or she had already completed them prior to interacting with the virtual machine). As before, the experimenter would step out of the room while the participant completed the questionnaire(s).

Finally, the experimenter would return and debrief the participant. At this point, the purpose of the experiment, and the nature of the virtual machine the participant interacted with, was explained in full by the experimenter. The puppetmaster was also introduced to the participant at this point. After providing a written copy of the debriefing

materials (Appendix C.2), the participants were thanked by the experimental team, and their SONA participation credit was assigned.

3.3 Results

As mentioned, this is both an exploratory study and a proof-of-concept of the usefulness of Turing tests as reflexive experimental apparatus. Therefore, I wish to take care not to miss any interesting results that, while not statistically significant, might nonetheless serve to inform future research questions. Therefore, when testing hypotheses in this study, I have assigned a level of statistical significance at $p < .05$. However, I have also assigned what I shall call an “exploratory” level of significance at $p < .1$. I shall take care not to draw any firm conclusions with tests that return p -values less than the exploratory p -value given here. However, I will offer some speculation on how these observations might inform future research. Now, on to the study.

3.3.1 Sample Characteristics

An initial analysis of some of the questionnaire data that were collected for this experiment, with a view toward gauging our participants’ prior experience with virtual reality, sociable robots, and other similar technologies, is presented below. As shown in Table 1, a majority of the participants reported having used a virtual reality device, such as the Oculus Rift, prior to participating in this study. However, only one participant reported having used a social VR application such as AltspaceVR. Likewise, only one participant reported having interacted with a humanoid robot prior to this experiment. Therefore, even though more than half of the participants were familiar with VR, a larger

majority of them reported no experience interacting with the kinds of machines that they interacted with during the experimental tasks described herein.

Table 1: Experience with Virtual Reality and Robots

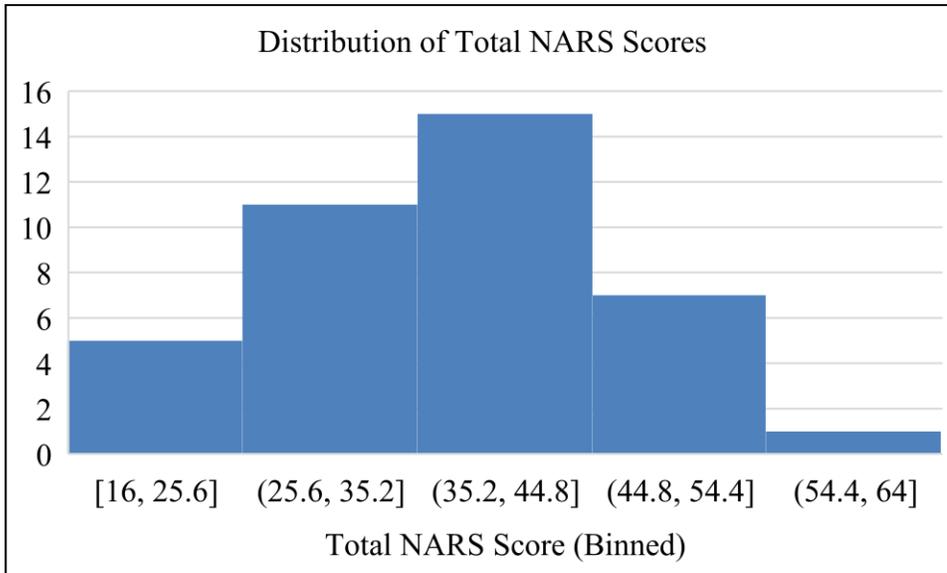
Questionnaire Item	Response	
	Yes	No
Have you ever used a virtual reality device?	25	14
Have you ever used a social VR application?	1	38
Have you ever used an augmented reality application?	3	36
Have you ever interacted with a humanoid robot?	1	38

Table 2 (below) illustrates the participants' average scores on each of the three subscales of the Negative Attitudes Toward Robots Scale, as well as the participant pool's total average NARS score. Figure 7 (below) illustrates the distribution of the participants' total average NARS scores.

Table 2: Total Mean NARS Scores

	N	Minimum	Maximum	Mean	Std. Deviation
S1: Interactions with Robots	39	6	26	13.69	4.572
S2: Social Influence of Robots	39	5	22	14.64	4.398
S3: Emotional Interactions with Robots	39	3	14	8.69	2.330
Total NARS Score	39	16	61	37.03	9.295
Valid N (listwise)	39				

Figure 7: Distribution of Total NARS Scores

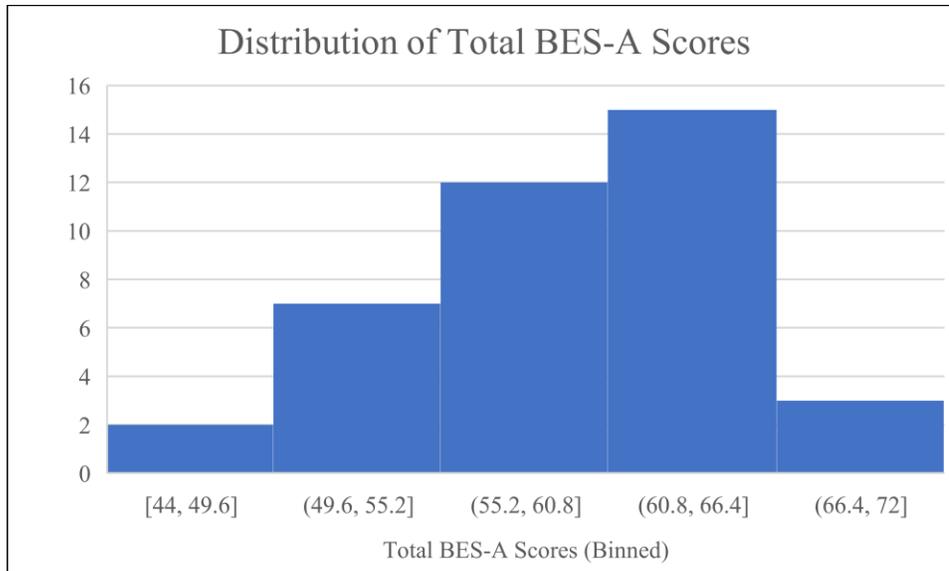


Group measures of dispositional empathy using the two-factor model of the Basic Empathy Scale in Adults are given in Table 3 (below), including the participant pool's average scores on the Affective Empathy subscale and the Cognitive Empathy subscale, as well as the total average BES-A score for the group. Distributions of the participant pool's total average BES-A scores are illustrated below (Figure 8).

Table 3: Total Mean BES-A Scores

	N	Minimum	Maximum	Mean	Std. Deviation
Total Affective Score	39	20	37	30.36	4.101
Total Cognitive Score	39	23	35	29.08	2.813
Total BES-A Score	39	44	72	59.44	5.394
Valid N (listwise)	39				

Figure 8: Distribution of Total BES-A Scores



An independent samples T-test did not reveal any significant differences between experimental conditions with respect to the distribution of the sample group's mean total NARS scores ($t(37)=.912$, $p=.367$). Similarly, no significant differences were found between the two experimental conditions with respect to the group's mean total BES-A scores ($t(37)=-.303$, $p=.763$). We are therefore confident that our assignment of participants into each experimental condition was sufficiently random.

Next, the participants' answers to selected questionnaire items were considered. Recall that the participants indicated the degree to which they agreed or disagreed with several statements describing their experiences interacting with one of the two virtual machines, depending upon the condition to which they were assigned. They did so by rating whether they agreed or disagreed with the presented statements using Likert-type scales (1=Strongly Disagree, 2=Disagree, 3=Neither Agree nor Disagree, 4 =Agree, 5

=Strongly Agree.). The participants were also asked to indicate, on a scale of 1-10, how immersed they felt during their experience. A Shapiro-Wilk test of normality indicated that the distribution of scores for each of the questionnaire items considered below deviated from normality ($p < .001$). Likewise, the participants' ratings of their sense of immersion also differed from normality ($p < .05$). A Mann-Whitney U test did not reveal any significant differences between the two experimental groups' sense of immersion, nor between their reported experiences interacting with the virtual machine. Group means for the participants' responses to these statements, together with experimental effects, are given below (Table 4).

Table 4: Group Means – Experience with the Virtual Agent

Experimental Condition		N	Mean	Std. Deviation	Std. Error	Experimental Effects (Mann-Whitney U)
On a scale of 1-10, please rate how immersed in the virtual environment you felt during this experiment.	Robotic	15	7.80	1.320	.341	$U=100.5, p=.608$
	Human	15	8.00	1.195	.309	
I liked the virtual character.	Robotic	21	4.10	.889	.194	$U=166.5, p=.503$
	Human	18	3.83	1.043	.246	
Interacting with the virtual character was fun.	Robotic	21	4.24	.889	.194	$U=146.0, p=.173$
	Human	18	4.61	.608	.143	
Interacting with the virtual character was frustrating.	Robotic	21	1.71	.845	.184	$U=144.0, p=.144$
	Human	18	1.33	.594	.140	
Interacting with the virtual character was annoying.	Robotic	21	1.43	.676	.148	$U=154.5, p=.202$
	Human	18	1.17	.383	.090	

3.3.2 Experimental Results

First, I consider the **(b)**-type questions of whether the appearance and behaviour of the machine plays any role in people's perceptions of the machine's capacity for agency and experience, and, upon the participants' judgements about what kind of reasoning agent controls the machine. To facilitate this analysis a principal components analysis was conducted on the perceived experience questionnaire items ("It seems like this character is able to experience emotions," "It seems like this character is able to experience physical sensations," and "It seems like this character has conscious experience") and on the perceived agency questions ("It seems like this character is able to control how it moves its body," "It seems like this character can interact with its environment," and "It seems like this character can control what it thinks and speaks").

Recall that each of the participants had to rate the extent to which they agreed/disagreed with each questionnaire item using a 5-point Likert-type scale. Let's begin with participants' answers to the perceived experience questions, the mean ratings of which are given in the table 5 (below).

A KMO test revealed a sampling adequacy of .695. The principal components analysis identified one component (Eigenvalue=2.154) that explained 71.81 percent of the variance in the participants' answers to the perceived capacity for experience questions. As shown in the component matrix below (table 6) the factor that loaded the highest to this component was "It seems like this character has conscious experience." The other two factors, however, still loaded very highly onto this component. Therefore, we extracted a new factor from these three components, which was called "Perceived Experience."

Table 5: Mean Perceived Experience Ratings

Questionnaire Item	Mean	Std. Dev.	N
It seems like this character is able to experience emotions.	3.08	1.244	39
It seems like this character is able to experience physical sensations.	1.90	1.021	39
It seems like this character has conscious experience.	2.87	1.151	39

Table 6: Component Loadings for Perceived Experience

	Component Loadings
It seems like this character is able to experience emotions.	0.840
It seems like this character is able to experience physical sensations.	0.820
It seems like this character has conscious experience.	0.881

Table 7: Mean Perceived Agency Ratings

Questionnaire Items	Mean	Std. Dev.	N
It seems like this character is able to control how it moves its body	3.05	1.255	39
It seems like this character can interact with its environment.	3.03	1.158	39
It seems like this character can control what it thinks and speaks.	3.67	1.034	39

Table 8: Component Loadings for Perceived Agency

	Component Loadings
It seems like this character is able to control how it moves its body.	0.571
It seems like this character can interact with its environment	0.856
It seems like this character can control what it thinks and speaks.	0.784

Next, this process was repeated for the perceived agency questionnaire items. Mean ratings for the perceived agency questions are given in the table 7 (above). A KMO test of sampling adequacy returned a value of 0.546, somewhat lower than the sampling adequacy of the perceived capacity for experience questions. A principal components analysis identified one component (Eigenvalue=1.673) that explained 55.778 percent of the variance in the participants' answers to the perceived capacity for agency questions. As shown in the component matrix in table 8 (above), the factor that loaded the highest to this component was "It seems like this character can interact with its environment." The other two components, however, still loaded reasonably well onto this component. Therefore, we extracted a new factor from these three components, which was called "Perceived Agency."

Table 9: Group Means for Perceived Experience and Perceived Agency

		N	Mean	Std. Deviation	Std. Error Mean	Experimental Effects (Independent T-test)
Perceived Experience	Robotic	21	-.0407055	1.06026934	.23136974	$t(37)=-0.271$, $p=0.788$
	Humanoid	18	.0474898	.95305045	.22463615	
Perceived Agency	Robotic	21	-.0775777	1.06122716	.23157875	$t(37)=-0.518$, $p=0.607$
	Humanoid	18	.0905073	.94564345	.22289030	

The new factors of Perceived Experience and Perceived Agency were standardized, such that scores above zero were considered "high" and scores below zero were considered "low" with respect to the participants' perceptions of the capacities of experience and agency of the virtual character. The means for the new standardized scores for participants' ratings of the agents' perceived capacities for experience and for

agency are given in the table 9 (above). A Shapiro-Wilk test of normality returned a non-significant result when the normality of the distribution of the above-given scores was tested. However, as is shown in the above table, an Independent Samples T-test did not reveal any significant between-group differences in the mean standardized perceived agency and experience scores of the virtual agents.

Concerning the **(b)**-type question of whether the machines appearance and behaviour – which in this case was either stereotypically robotic or more human-like – plays any role in people’s perceptions of agency and experience, these results indicate that we cannot reject the null hypothesis (H_{50} : The appearance and behaviour of the machine plays no role in people’s perceptions of the machine’s capacities for agency and experience).

Next, we considered the participants’ ratings of their levels of agreement/disagreement with the “Turing test questions,” which were designed in order to see what kind of reasoning agent – a person, or an AI – was controlling the virtual agent that they interacted with. Here, the participants were shown two different statements, one at a time, and asked to indicate the degree with which they agreed or disagreed with them:

Questionnaire Item: “The character was probably controlled by an artificial intelligence.”

Questionnaire Item: “The character was probably controlled by a person.”

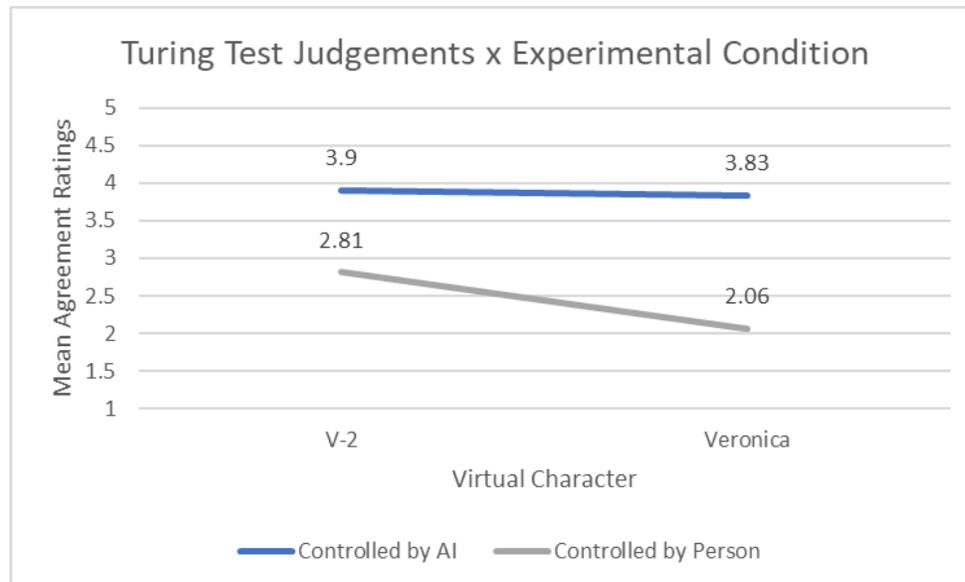
The participants provided their ratings so by indicating the degree with which they agreed or disagreed with the statements using 5-point Likert-type scales. A Shapiro-Wilk test

revealed that the distribution of the participants' responses differed from normality with respect to both of these questions ($p < .001$). Group means for the participants' ratings and experimental effects are given below (table 10).

Table 10: Group Means – Turing Test Questions

Experimental Condition		N	Mean	Std. Deviation	Std. Error Mean	Experimental Effects (Mann- Whitney U test)
The character was probably controlled by an artificial intelligence.	Robotic	21	3.90	1.044	.228	$U=176.5, p=.709$
	Humanoid	18	3.83	.924	.218	
The character was probably controlled by a person.	Robotic	21	2.81	1.327	.290	$U=126.5, p=.068$
	Humanoid	18	2.06	1.211	.286	

Figure 9: Participants' Turing Test Judgements



A Mann-Whitney U test did not reveal any significant differences in terms of agree/disagreement ratings with respect to the statement “The character was probably

controlled by an artificial intelligence.” However, an exploratorily significant difference was observed between the agreement/disagreement ratings made by the participants in the humanoid condition ($M=2.06$, $SD=1.211$) and the robotic condition ($M=2.81$, $SD=1.327$) with respect to the statement “The character was probably controlled by a person” ($U=126.5$, $p=.068$). The differences between the mean ratings on each of these judgements is illustrated above in figure 9.

The differences in the mean ratings of agreement/disagreement with the two Turing-test questionnaire items was not found to be statistically significant. I therefore cannot reject the null hypothesis (H_0 : The appearance and behaviour of the machine plays no role in people’s judgements about what kind of reasoning agent controls the machine). However, the mean difference was significant at $p<0.1$, which was set as an exploratory p -value for this experiment. Perhaps with a larger sample size or more pronounced differences between the virtual machines, the results would have been different. But in any case, I will return to the matter of how to understand these results in the discussion section.

Next, I consider the hypotheses related to **(a)**-type questions, as described above (section 3.1). The first of these research questions, as we have seen, is: do people’s levels of dispositional cognitive and affective empathy play some role when it comes to making judgements about the kind of machines that they are interacting with in Turing tests? The second **(a)**-type question is whether people’s levels of dispositional empathy play any role in their perception of the agency and experience of the machines they are interacting with in Turing tests. In order to facilitate the analysis with a view to addressing these questions, it was decided to reorganize our participant pool into high/low cognitive

empathy groups, and high/low affective empathy groups. Let's begin with cognitive empathy. Our participants' scores on this subscale of the BES-A were divided into five groups, where each group of participants' scores was assigned a value of 1-5 based upon the cut-points provided in table 11 (below). This process was repeated for the affective empathy subscale scores of the BES-A questionnaire items that our participants answered. Here, the participant pool was again divided into five groups. Each participants' score could be assigned a value of 1-5 based upon the cut-points shown in table 11.

We then reorganized our participants' cognitive empathy subscale scores into two groups. Participants whose scores were 3 were excluded altogether, leaving us with a low cognitive empathy group (made up of those with scores of 1 and 2) and a high cognitive empathy group (made up of scores 4 and 5). We assigned the low cognitive empathy group a value of 1.00 ($N=16$) and the high cognitive empathy group a value of 2.00 ($N=14$). Note nine of the participants were excluded, falling neither in the low or high cognitive empathy group, reducing the total participant pool for the cognitive empathy group ($N=30$). The participants' affective empathy subscale scores were also reorganized into two groups. As before, participants whose scores were 3 were excluded altogether, leaving us with a low affective empathy group (made up of scores 1 and 2) and an affective empathy group (made up of scores 4 and 5). We assigned the low affective empathy group a value of 1.00 ($N=17$) and the high affective empathy group a value of 2.00 ($N=13$). Once again, nine of the participants were excluded, falling neither in the low or high affective empathy group, reducing the total participant pool for this group ($N=30$).

Table 11: Cut-Points for Cognitive and Affective BES-A Scores

Cognitive Empathy Cut-Points		
	Value	Label
1	27.00	<=27
2	28.00	28-28
3	30.00	29-30
4	32.00	31-32
5	HIGH	33+
Affective Empathy Cut-Points		
	Value	Label
1	26.00	<=26
2	30.00	27-30
3	32.00	31-32
4	34.00	33-34
5	HIGH	34+

Having reorganized the data to facilitate further analysis, it was decided to first examine whether our participants' levels of dispositional cognitive empathy played any role in their perceptions of the virtual machines' capacities for agency and experience. As is shown in the table below, a One-Way ANOVA did not reveal any significant main effects of dispositional cognitive empathy on either the perceived capacity for experience (table 13) nor the perceived capacity for agency (see table 15). Further, a Bayesian ANOVA returned a Bayes Factor in support of the null hypothesis with respect to perceived experience (see table 14 below) and perceived agency (see table 17 below).

Table 12: Mean Perceived Experience Ratings of Low & High Cognitive Empathy Groups

Experimental Condition	Cog. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	-.0735862	1.24770045	8
	2.00	.0914714	.99744194	8
	Total	.0089426	1.09454721	16
Humanoid	1.00	.2682699	1.17726457	8
	2.00	.0399554	.74897619	6
	Total	.1704209	.98781860	14
Total	1.00	.0973419	1.18508566	16
	2.00	.0693931	.86727507	14
	Total	.0842991	1.03141034	30

Table 13: Tests of Between-Subject Effects – Dependent Variable: Perceived Experience

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	.482 ^a	3	.161	.138	.937	.016	.413	.072
Intercept	.196	1	.196	.168	.685	.006	.168	.068
Exp. Condition	.156	1	.156	.133	.718	.005	.133	.064
Cog. Emp. (Low/High)	.007	1	.007	.006	.937	.000	.006	.051
Exp. Condition* Cog. Emp. (Low/High)	.286	1	.286	.245	.625	.009	.245	.076
Error	30.368	26	1.168					
Total	31.064	30						
Corrected Total	30.850	29						

a. R Squared = .016 (Adjusted R Squared = -.098)

b. Computed using alpha = .05

Table 14: Bayesian ANOVA – Model Comparison: Perceived Experience

Models	P(M)	P(M data)	BF _M	BF ₁₀	error %
Null model	0.200	0.526	4.446	1.000	
Cog Emp (Low/High)	0.200	0.182	0.890	0.346	2.060e-5
Exp. Condition	0.200	0.194	0.964	0.369	1.453e-4
Cog. Emp. (Low/High) + Exp. Condition	0.200	0.067	0.285	0.126	2.393
Cog. Emp. (Low/High) + Exp. Condition + Cog. Emp. (Low/High) * Exp. Condition	0.200	0.031	0.127	0.059	2.261

Table 15: Mean Perceived Agency Ratings of Low & High Cognitive Empathy Groups

Experimental Condition	Cog. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	-.1446887	.94243586	8
	2.00	.3352065	1.22532064	8
	Total	.0952589	1.08469257	16
Humanoid	1.00	-.1305456	.96189760	8
	2.00	.5277270	.94372251	6
	Total	.1515712	.97725945	14
Total	1.00	-.1376172	.91995748	16
	2.00	.4177153	1.07739046	14
	Total	.1215380	1.01857770	30

Table 16: Tests of Between-Subject Effects – Dependent Variable: Perceived Agency

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	2.431 ^a	3	.810	.762	.526	.081	2.285	.190
Intercept	.638	1	.638	.599	.446	.023	.599	.116
Exp. Condition	.079	1	.079	.074	.788	.003	.074	.058
Cog. Emp. (Low/High)	2.392	1	2.392	2.248	.146	.080	2.248	.303
Exp. Condition * Cog. Emp. (Low/High)	.059	1	.059	.055	.816	.002	.055	.056
Error	27.657	26	1.064					
Total	30.531	30						
Corrected Total	30.088	29						

a. R Squared = .081 (Adjusted R Squared = -.025)

b. Computed using alpha = .05

Table 17: Bayesian ANOVA – Model Comparison: Perceived Agency

Models	P(M)	P(M data)	BF _M	BF ₁₀	error %
Null model	0.200	0.388	2.531	1.000	
Cog. Emp. (Low/High)	0.200	0.317	1.856	0.818	0.005
Exp. Condition	0.200	0.135	0.623	0.348	2.727e-5
Cog. Emp. (Low/High) + Exp. Condition	0.200	0.112	0.502	0.288	0.827
Cog. Emp. (Low/High) + Exp. Condition + Cog. Emp. (Low/High) * Exp. Condition	0.200	0.049	0.207	0.127	1.097

Next, it was decided to examine whether our participants' self-reported levels of dispositional affective empathy played any role in their perceptions of the virtual machines' capacities for agency and experience. A One-Way ANOVA did not indicate any significant main effects of dispositional affective empathy on either the perceived capacity for experience (table 19, below) nor the perceived capacity for agency (table 22, below). Likewise, a Bayesian ANOVA returned a Bayes Factor in support of the null hypothesis in the case of perceived experience (table 20, below), and perceived agency (table 23, below).

Table 18: Mean Perceived Experience Ratings of Low & High Affective Empathy Groups

Experimental Condition	Aff. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	-.3821493	1.25188597	11
	2.00	.3023365	.74976663	6
	Total	-.1405661	1.12644069	17
Humanoid	1.00	-.0309871	1.18706292	6
	2.00	.4603980	.65317966	7
	Total	.2336049	.93030250	13
Total	1.00	-.2582097	1.20406937	17
	2.00	.3874465	.67400097	13
	Total	.0215747	1.04582608	30

Table 19: Tests of Between-Subject Effects – Dependent Variable: Perceived Experience

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	3.630 ^a	3	1.210	1.120	.359	.114	3.361	.266
Intercept	.216	1	.216	.199	.659	.008	.199	.071
Exp. Condition	.457	1	.457	.423	.521	.016	.423	.096
Aff. Emp. (Low/High)	2.438	1	2.438	2.257	.145	.080	2.257	.304
Exp. Condition* Aff. Emp. (Low/High)	.066	1	.066	.061	.807	.002	.061	.056
Error	28.088	26	1.080					
Total	31.733	30						
Corrected Total	31.719	29						

a. R Squared = .114 (Adjusted R Squared = .012)

b. Computed using alpha = .05

Table 20: Bayesian ANOVA – Model Comparison: Perceived Experience

Models	P(M)	P(M data)	BF _M	BF ₁₀	error %
Null model	0.200	0.314	1.834	1.000	
Aff. Emp. (Low/High)	0.200	0.329	1.965	1.048	0.004
Exp. Condition	0.200	0.155	0.735	0.494	$2.156e^{-4}$
Aff. Emp. (Low/High) + Exp. Condition	0.200	0.140	0.652	0.446	1.360
Aff. Emp. (Low/High) + Exp. Condition + Aff. Emp. (Low/High) * Exp. Condition	0.200	0.061	0.259	0.194	1.530

Table 21: Mean Perceived Agency Ratings of Low & High Affective Empathy Groups

Experimental Condition	Aff. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	-.3334803	1.09476004	11
	2.00	-.0004874	.76939305	6
	Total	-.2159534	.98028425	17
Humanoid	1.00	.1505277	.86492921	6
	2.00	.4475088	1.12403580	7
	Total	.3104406	.98345472	13
Total	1.00	-.1626539	1.01965075	17
	2.00	.2407413	.96561659	13
	Total	.0121507	1.00039213	30

Table 22: Tests of Between Subject Effects – Dependent Variable: Perceived Agency

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	2.757 ^a	3	.919	.910	.450	.095	2.729	.221
Intercept	.123	1	.123	.122	.730	.005	.122	.063
Exp. Condition	1.532	1	1.532	1.516	.229	.055	1.516	.220
Aff. Emp. (Low/High)	.700	1	.700	.693	.413	.026	.693	.126
Exp. Condition * Aff. Emp. (Low/High)	.002	1	.002	.002	.962	.000	.002	.050
Error	26.266	26	1.010					
Total	29.027	30						
Corrected Total	29.023	29						

a. R Squared = .095 (Adjusted R Squared = -.009)

b. Computed using alpha = .05

Table 23: Bayesian ANOVA – Model Comparison: Perceived Agency

Models	P(M)	P(M data)	BF _M	BF ₁₀	error %
Null model	0.200	0.353	2.180	1.000	
Aff. Emp. (Low/High)	0.200	0.192	0.953	0.545	1.828e-4
Exp. Condition	0.200	0.269	1.471	0.762	0.005
Aff. Emp. (Low/High) + Exp. Condition	0.200	0.131	0.601	0.370	0.795
Aff. Emp. (Low/High) + Exp. Condition + Aff. Emp. (Low/High) * Exp. Condition	0.200	0.055	0.235	0.157	1.881

So, when it comes to the question of whether high vs. low levels of self-reported dispositional cognitive empathy plays a role in people’s perceptions of the agents’ capacities for agency and experience, these results mean that I cannot reject the null

hypotheses that pertains to these research questions (H4₀: People’s levels of dispositional cognitive empathy don’t play a role in their perceptions of agency and experience; and H3₀: People’s levels of dispositional affective empathy don’t play a role in their perceptions of agency and experience).

Next, it was decided to re-examine the **(a)**-type questions with respect to high vs. low dispositional affective and cognitive empathy. Let us begin with the questionnaire item “The character was probably controlled by a person.” As shown below, a One-Way ANOVA failed to return any significant main effect of affective empathy (table 25, below), or of cognitive empathy (table 28, below), of the participants ratings of agreement/disagreement with this questionnaire item. Similarly, a Bayesian ANOVA returned a Bayes factor in favor of the null hypothesis with respect to affective empathy (table 26, below) and to cognitive empathy (table 29, below).

Table 24: Low & High Affective Empathy Group Mean Agreement Ratings for “The character was probably controlled by a person.”

Experimental Condition	Aff. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	2.91	1.375	11
	2.00	2.33	1.366	6
	Total	2.71	1.359	17
Humanoid	1.00	2.17	1.472	6
	2.00	2.14	1.345	7
	Total	2.15	1.345	13
Total	1.00	2.65	1.412	17
	2.00	2.23	1.301	13
	Total	2.47	1.358	30

Table 25: Tests of Between Subjects Effects – Dependent Variable: “The character was probably controlled by a person.”

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	3.534 ^a	3	1.178	.613	.612	.066	1.840	.160
Intercept	160.888	1	160.888	83.774	.000	.763	83.774	1.000
Exp. Condition	1.535	1	1.535	.799	.380	.030	.799	.138
Aff. Emp. (Low/High)	.634	1	.634	.330	.571	.013	.330	.086
Exp. Condition * Aff. Emp. (Low/High)	.537	1	.537	.280	.601	.011	.280	.080
Error	49.933	26	1.920					
Total	236.000	30						
Corrected Total	53.467	29						

a. R Squared = .066 (Adjusted R Squared = -.042)

b. Computed using alpha = .05

Table 26: Bayesian ANOVA – Model Comparison: “The character was probably controlled by a person.”

Models	P(M)	P(M data)	BF _M	BF ₁₀	error %
Null model	0.200	0.427	2.983	1.000	
Aff. Emp. (Low/High)	0.200	0.192	0.949	0.449	5.303e-4
Exp. Condition	0.200	0.235	1.227	0.549	2.397e-4
Aff. Emp. (Low/High) + Exp. Condition	0.200	0.098	0.433	0.228	0.860
Aff. Emp. (Low/High) + Exp. Condition + Aff. Emp. (Low/High) * Exp. Condition	0.200	0.049	0.205	0.114	3.651

Table 27: Low & High Cognitive Empathy Group Mean Ratings for “The character was probably controlled by a person.”

Experimental Condition	Cog. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	3.38	1.302	8
	2.00	2.63	1.408	8
	Total	3.00	1.366	16
Humanoid	1.00	2.38	1.188	8
	2.00	2.17	1.472	6
	Total	2.29	1.267	14
Total	1.00	2.88	1.310	16
	2.00	2.43	1.399	14
	Total	2.67	1.348	30

Table 28: Tests of Between-Subjects Effects – Dependent Variable: “The character was probably controlled by a person.”

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	6.208 ^a	3	2.069	1.158	.345	.118	3.474	.275
Intercept	205.157	1	205.157	114.814	.000	.815	114.814	1.000
Exp. Condition	3.926	1	3.926	2.197	.150	.078	2.197	.298
Cog. Emp. (Low/High)	1.696	1	1.696	.949	.339	.035	.949	.155
Exp. Condition * Cog. Emp. (Low/High)	.542	1	.542	.303	.587	.012	.303	.083
Error	46.458	26	1.787					
Total	266.000	30						
Corrected Total	52.667	29						

a. R Squared = .118 (Adjusted R Squared = .016)

b. Computed using alpha = .05

Table 29: Bayesian ANOVA – Model Comparison – “The character was probably controlled by a person.”

Models	P(M)	P(M data)	BF _M	BF ₁₀	error %
Null model	0.200	0.352	2.178	1.000	
Cog. Emp. (Low/High)	0.200	0.165	0.793	0.469	4.821e-4
Exp. Condition	0.200	0.274	1.511	0.778	0.005
Cog. Emp. (Low/High) + Exp. Condition	0.200	0.140	0.649	0.396	0.766
Cog. Emp. (Low/High) + Exp. Condition + Cog. Emp. (Low/High) * Exp. Condition	0.200	0.068	0.293	0.193	2.497

Next, the participants’ ratings of agreement with the questionnaire item “The character was probably controlled by an artificial intelligence” were examined. The results of a One-Way ANOVA did not reveal any significant main effects of self-reported levels of dispositional affective empathy on the participants’ ratings of agreement with this statement, as shown in the table 31 below. Likewise, a Bayesian ANOVA returned a Bayes factor in favor of the null hypothesis (see table 32).

Table 30: Low & High Affective Empathy Group Mean Ratings for “The character was probably controlled by an artificial intelligence.”

Experimental Condition	Aff. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	3.91	1.136	11
	2.00	4.33	.816	6
	Total	4.06	1.029	17
Humanoid	1.00	3.67	1.211	6
	2.00	4.14	.690	7
	Total	3.92	.954	13
Total	1.00	3.82	1.131	17
	2.00	4.23	.725	13
	Total	4.00	.983	30

Table 31: Tests of Between Subjects Effects – Dependent Variable: “The character was probably controlled by an artificial intelligence.”

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	1.567 ^a	3	.522	.514	.676	.056	1.541	.140
Intercept	454.356	1	454.356	446.915	.000	.945	446.915	1.000
Exp. Condition	.330	1	.330	.325	.573	.012	.325	.085
Aff. Emp. (Low/High)	1.430	1	1.430	1.406	.246	.051	1.406	.208
Exp. Condition * Aff. Emp. (Low/High)	.005	1	.005	.005	.946	.000	.005	.050
Error	26.433	26	1.017					
Total	508.000	30						
Corrected Total	28.000	29						

a. R Squared = .056 (Adjusted R Squared = -.053)

b. Computed using alpha = .05

Table 32: Bayesian ANOVA – Model Comparison – “The character was probably controlled by an artificial intelligence.”

Models	P(M)	P(M data)	BF _M	BF ₁₀	error %
Null model	0.200	0.445	3.202	1.000	
Aff. Emp. (Low/High)	0.200	0.249	1.325	0.560	4.016e-4
Exp. Condition	0.200	0.162	0.775	0.365	9.549e-5
Aff. Emp. (Low/High) + Exp. Condition	0.200	0.101	0.448	0.227	5.513
Aff. Emp. (Low/High) + Exp. Condition + Aff. Emp. (Low/High) * Exp. Condition	0.200	0.044	0.182	0.098	1.108

However, as shown in table 34 (below), when it comes to the participants’ self-reported levels of dispositional cognitive empathy, a significant main effect of cognitive empathy

(low vs. high) was found ($F(1,29)=5.016, p=0.034$). Additionally, as illustrated in table 35 (below), a Bayesian ANOVA returned a marginal Bayes Factor in favour of the alternative hypothesis ($BF_{10}=2.597$). In other words, this suggests that the results observed here are approximately 2.6 times more likely to occur under a model that includes an effect of high/low dispositional cognitive empathy than under a model that doesn't.

Table 33: Low & High Cognitive Empathy Group Mean Agreement Ratings for “The character was probably controlled by an artificial intelligence.”

Experimental Condition	Cog. Emp. (Low/High)	Mean	Std. Deviation	N
Robotic	1.00	3.38	1.061	8
	2.00	4.37	1.061	8
	Total	3.88	1.147	16
Humanoid	1.00	3.50	1.069	8
	2.00	4.17	.753	6
	Total	3.79	.975	14
Total	1.00	3.44	1.031	16
	2.00	4.29	.914	14
	Total	3.83	1.053	30

Table 34: Tests of Between-Subjects Effects – Dependent Variable: “The character was probably controlled by an artificial intelligence.”

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power ^b
Corrected Model	5.583 ^a	3	1.861	1.820	.168	.174	5.461	.417
Intercept	438.782	1	438.782	429.154	.000	.943	429.154	1.000
Exp. Condition	.013	1	.013	.013	.912	.000	.013	.051
Cog. Emp. (Low/High)	5.128	1	5.128	5.016	.034	.162	5.016	.578
Exp. Condition * Cog. Emp. (Low/High)	.205	1	.205	.201	.658	.008	.201	.072
Error	26.583	26	1.022					
Total	473.000	30						
Corrected Total	32.167	29						

a. R Squared = .174 (Adjusted R Squared = .078)

b. Computed using alpha = .05

Table 35: Bayesian ANOVA – Model Comparison: “The character was probably controlled by an artificial intelligence.”

Models	P(M)	P(M data)	BF_M	BF₁₀	error %
Null model	0.200	0.188	0.928	1.000	
Cog. Emp. (Low/High)	0.200	0.489	3.828	2.597	8.827e - 4
Exp. Condition	0.200	0.066	0.284	0.352	4.205e - 5
Cog. Emp. (Low/High) + Exp. Condition	0.200	0.168	0.810	0.894	0.773
Cog. Emp. (Low/High) + Exp. Condition + Cog. Emp. (Low/High) * Exp. Condition	0.200	0.088	0.386	0.468	11.548

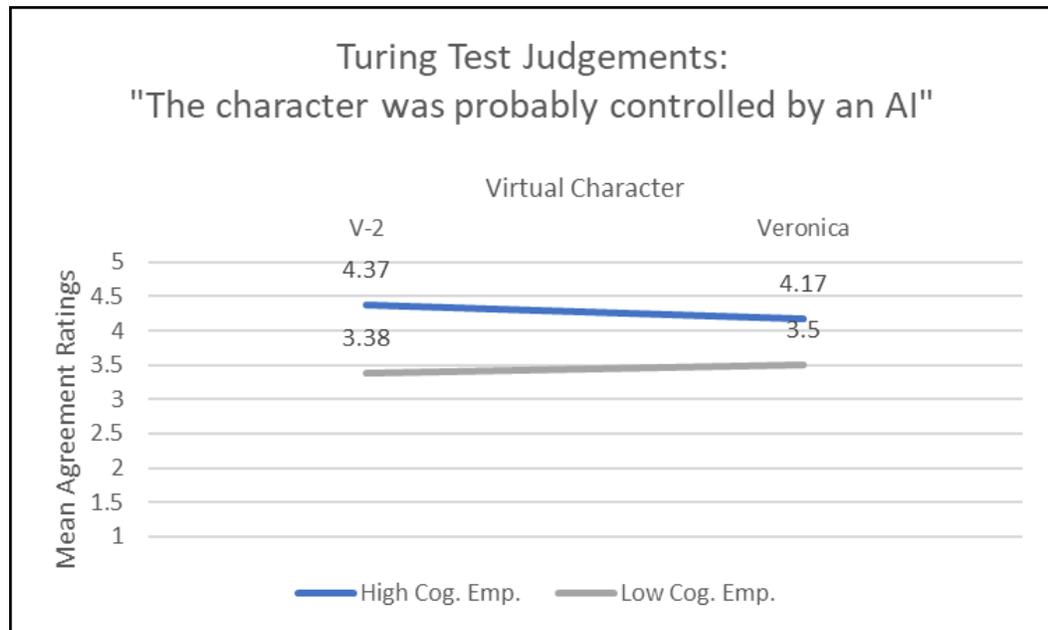
Regarding the question of whether differences in people’s self-reported levels of dispositional affective empathy plays some role in the judgements that they make about the kind of reasoning agent that controls the character with which they interacted, we cannot reject the null hypothesis (H1₀: People’s levels of dispositional affective empathy don’t play a role in such judgements.) Now, recall that the hypotheses related to the question of whether differences in people’s self-reported levels of dispositional cognitive empathy play a role in the judgements they make about the kind of reasoning agent that controls the character with which they interacted are:

H2₀: People’s levels of dispositional cognitive empathy don’t play a role in such judgements.

H2₁: People’s levels of dispositional cognitive empathy do play a role in such judgements.

The results outlined above indicate that we can reject H2₀. Rather, the results indicate that people whose dispositional cognitive empathy placed them in the “high” group provided significantly different ratings of agreement with this questionnaire item than did those in the “low” group, regardless of experimental condition.

Figure 10: Low and High Cognitive Empathy Group Mean Agreement Ratings: “The character was probably controlled by an artificial intelligence.”



More specifically, and as illustrated in figure 10 (above), the ratings of agreement provided by those in the high dispositional cognitive empathy group (Robotic: $M=4.37$, $SD=1.061$; Humanoid: $M=4.17$, $SD=0.753$) are much closer to “strongly agree” (i.e. a 5 on the scale) with the questionnaire item than the ratings provided by those in the low dispositional cognitive empathy group (Robotic: $M=3.38$, $SD=1.061$; Humanoid: $M=3.5$, $SD=1.069$), which were closer to “uncertain”/”agree” (i.e. 3 and 4 on the scale). I shall discuss how I interpret these results in the following section.

6.4 Discussion

Before I discuss the results of this exploratory study, I must re-emphasize the fact that it is meant to be exploratory. For one of the principal aims of this study – aside from

serving as a proof-of-concept of my proposal that Turing tests are useful not just as conceptual prompts or as non-reflexive experimental apparatus, but also as reflexive experimental apparatus – is to explore whether variable empathic characteristics of people play a role in the judgements they make about the machines with which they interacted in this study. It is my hope that the findings I have presented above will help to formulate even more precise research questions – and more precise empirically testable hypotheses – which cognitive scientists can fruitfully address by deploying Turing tests as reflexive experimental apparatus.

That said, I think that there are a number of interesting findings that resulted from this exploratory study. For instance, with respect to our group of participants, the mean ratings they provided for their sense of immersion was not significantly different between the experimental conditions. That is, regardless of whether the participant interacted with V-2 or with Veronica, they seemed to be immersed in the experimental task. Similarly, no significant differences were found between experimental conditions with respect to whether the participants liked the virtual character, or whether they found interacting with it fun. Nor were any significant differences found with respect to whether they found interacting with the agent frustrating or annoying. Indeed, the average Likert-type ratings the participants provided seem to indicate that, regardless of the experimental condition to which they were assigned, they enjoyed interacting with the virtual agent, and did not find interacting with it frustrating or annoying. This is promising, for it suggests that Turing tests can be deployed as reflexive experimental apparatus in virtual environments without too much worry over whether the VR apparatus interferes with

immersion in the experimental task or makes Turing-test style interactions too difficult or frustrating for the user.

Another curious observation concerns the aforementioned **(b)**-type questions, and the participants' judgements with respect to the "Turing Test questions," i.e. the two statements that describe the virtual agent as probably being controlled by an AI and probably being controlled by a person. Upon hearing the description of the virtual machines used for this experiment, one might assume that by virtue of each agent's physical appearance, movement, and manner of speech, the participants would have largely agreed that V-2 was controlled by an AI but disagree that it was controlled by a person; likewise, that Veronica was controlled by a person, and disagree that she was controlled by an AI. In fact, these results show that participants in both experimental conditions agreed that the agent they interacted with was controlled by an AI. The unexpected finding, however, was that while the participants in the humanoid condition disagreed that Veronica was controlled by a person, the average Likert-type judgements made by participants assigned to the robotic condition are much closer to a rating of "uncertain" than to "disagree" with respect to whether V-2 was controlled by a person. Thus, an interesting incongruity emerged in the data with respect to the participants' Turing-test judgements, especially with respect to the above-described judgements of who or what was in control of V-2: a person, or an AI. Perhaps when it comes to people's judgements about which of these sorts of reasoning agents is in control of the machine, the appearance and behaviour of the machine matter. In other words, perhaps Turing (1950) missed the mark when he mused that "we should feel there was little point in trying to make a "thinking machine" more human by dressing it up in [...] artificial

flesh” (p. 434). However, as the results only achieved exploratory significance and not statistical significance here, I will not offer any firm conclusions about this at the present time. Further research will be required to address empirical questions surrounding this matter, but these are arguably questions that can, in principle, be addressed using Turing tests as reflexive experimental apparatus.

Arguably the most interesting – and indeed, the most surprising – result concerned the **(a)**-type questions addressed in this exploratory study. Although many of the statistical tests the data were subjected to returned non-significant results, something very interesting occurred when cognitive empathy was examined. Specifically, participants whose self-reported levels of dispositional cognitive empathy placed them in the high cognitive empathy group provided significantly different ratings of agreement to the statement “The character was probably controlled by an artificial intelligence” than did those whose self-reported dispositional cognitive empathy scores placed them in the low cognitive empathy group, regardless of the sort of machine they interacted with (more robot-like or more human-like). As discussed in the results section, the average ratings of agreement with this questionnaire item provided by the high cognitive empathy group were closer to “strongly agree” (i.e. a 5 on the Likert-type scale) than the ratings provided by those in the low dispositional cognitive empathy group, which were closer to “uncertain”/“agree” (i.e. 3-4 on the Likert-type scale). Although the initial sample size for this study was small ($N=39$), and the participant pool used for these tests was smaller ($N=30$), these results are certainly suggestive of something interesting that deserves further examination.

Specifically, these results might suggest that higher levels of dispositional cognitive empathy encourage people to judge that these machines were controlled by an artificial intelligence, and *not* by a person. This is especially interesting because the virtual machines used in this study *were* controlled by humans, as I specified earlier (section 3.2.2). For instance, consider that even though V-2 was created to seem obviously robotic, and Veronica was created to seem more human-like, both of these virtual machines fall well below T3 level indistinguishability.²⁴ Indeed, V-2 may have been *too* robotic for our purposes. For although Veronica was meant to seem much more human-like than V-2, many of the AI applications that our participants doubtlessly have interacted with previous to our experiment are arguably also more human-like than V-2, at least in terms of their verbal behaviour. So, Veronica might seem quite human-like when compared to V-2. Yet, when compared to Apple’s Siri or Microsoft’s Cortana, or any number of other virtual assistants, Veronica probably seems more AI-like than human-like.

Why, then, did the participants in the high dispositional cognitive empathy group make different judgements with respect to the statement “The character was probably controlled by an artificial intelligence” regardless of the experimental condition? Recall that Joliffe and Farrington (2006), the designers of the questionnaire used to collect these data – the BES-A – understand cognitive empathy as the ability to understand the emotions of another (as opposed to experiencing emotions that are consonant with another’s, i.e. affective empathy). But as I discussed in Chapter Two, following

²⁴ Recall that the important feature of the machines, for my purposes, was that the participants were still able to interact with them in a scenario where both the linguistic and robotic capacities of the machine were apparent to the participants.

Maibom's (2007; 2012; 2014) discussion of the various conceptions of empathy that are found throughout the literature, this is not the only way in which to understand cognitive empathy. In fact, I suspect that achieving cognitive empathy in the sense of understanding/comprehending the emotions of another probably employs the same mental machinery that underwrites theorizing or simulating in order to understand others' mental states in general. In this sense, cognitive empathy is just a specific application of mindreading, and mindreading may come to occur in several ways, e.g. by theorizing or by simulating. It may be that the measures of dispositional cognitive empathy that were collected here track not only people's ability to understand another's emotional experience, but at understanding other's mental experience in general (i.e. thoughts, beliefs, desires, *et cetera*). If so, then perhaps we can make sense of the results observed here.

Following this line of thinking, I speculate that those participants in the high cognitive empathy group, by virtue of their higher levels of dispositional cognitive empathy, may also possess higher dispositional mindreading abilities than those in the low cognitive empathy group. These participants might have judged that V-2 and Veronica were controlled by an AI with greater certainty than did those in the low cognitive empathy group because, on the one hand, they are good at working out what kinds of reasoning agents they are interacting with, and on the other, both of the virtual machines turned out to behave more like the aforementioned AIs that they have almost certainly interacted with before. The low cognitive empathy group, while arguably also familiar with such AIs, may have been less certain about what kind of reasoning agent controlled the machines owing to lower dispositional mindreading abilities.

Of course, this is far from a certain explanation of the results observed in this study. Then again, the aim of this exploratory study has been to offer a proof-of-concept of the usefulness of Turing tests as experimental apparatus, and to identify possible avenues of future research that pertain to the humans involved in human-machine interactions. I believe that this proof-of-concept has been largely successful, for an interesting finding that requires following up on was identified in this study. Moreover, I think that future research using Turing tests as I have done here, i.e. as reflexive experimental apparatus, can help shed light on: firstly, why people with various levels of dispositional cognitive empathy make various judgements about what sorts of reasoning agents control the machines with which they interact; and secondly, whether people's dispositional mindreading abilities (understood in the general sense) might also play a role in the judgements they make about the machines with which they interact. But whatever the answers to such questions turn out to be, I think my novel proposal that Turing tests are useful as reflexive experimental apparatus – that is, for answering empirical questions about the humans involved in human-machine interactions – has been effectively demonstrated.

Conclusion

In this investigation I have argued for, and attempted to illustrate, a novel understanding of the usefulness of Turing tests, namely, as reflexive experimental apparatus. That is, while Turing tests have (since Turing's introduction of the original Turing test) largely been seen as good for prompting philosophical reflection on the limits of our concept of the mental, or as good for addressing empirical questions about the machines involved in human-machine interactions, I have argued that they are also good for addressing empirical questions about the *humans* involved in human-machine interactions.

I completed the groundwork for this proposal in Chapter One, which concerned the two main understandings of the usefulness of Turing tests that are found in the relevant literature. I began by describing the Imitation Game, and I noted the relevant sense in which Turing tests are importantly similar to it, despite differing from it in various ways. I also clarified what Turing means by "machine" and "to think." I did so by showing that Turing's understanding of "machine" (in the relevant context of discussion) is limited to the digital computer. I also pointed out, following others, that Turing's understanding of thinking, or of the mental, is Cartesian in spirit. For on Turing's view, whether an entity can be said to think depends upon whether it possesses the capacity for domain-general reasoning, the capacity to express its reasoning using language, and the capacity to reason originally. Along the way, I supplied some additional motivational force for my novel understanding of the usefulness of Turing tests by discussing some of Turing's other contributions to the cognitive sciences, namely: the concepts he introduced that made possible computationalism, or the computational-representational theory of

mind. I argued that on computationalism, the question of whether machines can think – the questions Turing answers with the Imitation Game – is an easy question to answer without appealing to Turing tests.

I then illustrated the two main existing understandings of the usefulness of Turing tests. I showed that one of these understandings sees Turing tests as good for prompting philosophical reflection on the limits of our concept of the mental, i.e. as what I have called “conceptual prompts.” I also showed that the other understanding sees Turing tests as good for addressing empirical questions about the machines involved in human-machine interactions, i.e. as what I have called “non-reflexive experimental apparatus.” At the end of Chapter One, I provided a sketch of my novel understanding of the usefulness of Turing tests, which sees them as good for addressing empirical questions about the *humans* involved in human-machine interactions – as what I have called “reflexive experimental apparatus.” I further proposed that as reflexive experimental apparatus Turing tests may be fruitfully deployed to address the following two sorts of empirical questions about the humans so involved:

- (a) Are there variable characteristics of people that play some role in their judgements of whether machines are capable of thinking, possess mental states, emotions, and the like?
- (b) Do people’s judgements about these things also depend in some way upon how the machines look and behave?

In Chapter Two, I made my novel proposal (that Turing tests are as useful not just as conceptual prompts or non-reflexive experimental apparatus, but also as reflexive

experimental apparatus) more concrete with some further discussion about the sorts of machines, and the sorts of empirical questions about people, that I see Turing tests, deployed as reflexive experimental apparatus, as involving. Here, I drew upon the work of Steven Harnad. I argued that although the machines involved in the Turing tests may be situated at various levels of indistinguishability from human agents, Turing tests can still be fruitfully deployed as reflexive experimental apparatus even if the machines involved in the tests are situated at relatively low levels of indistinguishability from human beings. I then provided an analogy to clarify this proposal, which drew upon the work of android scientist Hiroshi Ishiguro. Ishiguro has used Turing tests to address empirical questions about the machines involved in human-machine interactions, where those machines are situated at lower levels of indistinguishability. I argued that, in a similar manner, researchers can also make use of Turing tests to address empirical questions about the humans involved in these interactions, even if machines are situated at low levels of indistinguishability.

Following this, I discussed the sorts of empirical questions about the humans involved in human-machine interactions that I think Turing tests, deployed as reflexive experimental apparatus, can help researchers to address. I argued that some good candidate research questions will involve the role of people's dispositional empathy, for there is a great deal of interest in concerning empathy in present human-robot interaction research. These research questions, more specifically, concerned the role that people's levels of dispositional empathy play in perceiving attributes of the mental, such as agency and experience, as well as the role it plays in people's judgements about what kind of reasoning agent is in control of the machine that is involved in the test.

In Chapter Three, I presented an exploratory study that was meant to serve as a “proof-of-concept” of my novel understanding of the usefulness of Turing tests, wherein I deployed a Turing test as reflexive experimental apparatus to address empirical questions of sorts **(a)** and **(b)**. More specifically, this study employed VR to address questions about the role that people’s levels of dispositional empathy play in the judgements they make about what kind of reasoning agent – an artificial intelligence, or a person – is in control of the machine with which they interact in a virtual environment. Here, I focused on an interesting observation, namely: participants with high levels of self-reported dispositional cognitive empathy made different judgements than did those with low levels of self-reported dispositional cognitive empathy with respect to the question of whether an artificial intelligence rather than a person was in control of the machine. That is, the participants who possess higher levels of dispositional cognitive empathy judged that the machines with which they interacted were controlled by an artificial intelligence with greater certainty than did those who possess lower levels of dispositional cognitive empathy.

I then suggested the possibility that these findings might be explained by the possibility that those who possess high dispositional cognitive empathy are, firstly, good at working out what kinds of reasoning agents they are interacting with owing to high dispositional mindreading abilities, and secondly, are familiar with examples of artificial intelligence that behave similar to the virtual machines used in this study, such as Apple’s Siri, Amazon’s Alexis, Google Home, or Microsoft’s Cortana. On the other hand, the people who possess low levels of dispositional cognitive empathy might therefore possess lower dispositional mindreading abilities, thereby causing them to make different

judgements with respect to whether an AI is in control of the machine with which they interacted despite their familiarity with the kinds of ubiquitous AIs mentioned above.

I therefore conclude that while a fully concrete explanation of these results is yet to be achieved, I have nonetheless succeeded at identifying avenues of future research that pertains to the mindreading capacities of people, and the role this might play in human-machine interaction. I further conclude that the proof-of-concept of the usefulness of Turing tests as reflexive experimental apparatus presented herein – which also constitutes an important part of my novel proposal concerning what Turing tests are good for – has been successful. One reason is that, in deploying Turing tests as reflexive experimental apparatus – the first instance of Turing’s test being deployed in such a way of which I am aware – I have achieved some interesting and unexpected findings with respect to questions of sorts **(a)** and **(b)** (but especially **(a)**-type questions). Another reason is that I think further research using Turing tests as reflexive experimental apparatus can shed additional light on these findings.

My intention in this thesis has not been to challenge the existing understanding of Turing tests as good for prompting philosophical reflection on the limits of our concept of the mental or for addressing empirical questions about the machines involved in human-machine interactions. Once again, my intended novel contribution has rather been to expand our understanding of the usefulness of Turing tests by showing that they are good for something else as well, namely, for addressing empirical questions about the humans involved in human-machine interactions. I submit that the results of my exploratory study in Chapter Three, combined with the foundational and conceptual work of Chapters One and Two, have fulfilled my intention.

Appendices

Appendix A - Virtual Agent Scripts

Note to the reader: V-2's and Veronica's dialogue is presented as normal text. Italicized text represents the anticipated responses of the participant. Text that appears in square brackets describes an event that occurs during the interaction scenario.

A.1 Robotic Agent Script

Introduction:

Hello, my name is V-2. I am a virtual sociable robot, created here in the Advanced Cognitive Engineering lab, to help understand how people interact with artificial agents. What is your name?

[Participants tells V-2 their name]

I am pleased to make your acquaintance. Would you like to play a game of twenty questions with me?

Yes.

Very good. Here are the rules. I can ask up to twenty questions about whatever it is you are thinking of. I can make a final guess at any time during the game, but if I guess wrong, or if I fail to guess in twenty questions or less, you win. Do you understand the rules?

Yes.

Excellent. And have you thought of something for me to guess?

Yes.

Understood. Then let us begin the game...

Inanimate Objects:

Airplane:

Here is my first question. Are you thinking of a living thing?

No.

Second question: are you thinking of a naturally occurring thing?

No.

Third question: are you thinking of an artificial thing?

Yes.

Fourth question: are you thinking of a machine of some kind?

Yes.

Fifth question: is the machine that you are thinking of a robot of some kind?

No.

Processing. Processing. Perhaps I should have asked a less specific question. Next question: would you say this machine is small?

No.

Seventh question: would you say that the machine you are thinking of is large?

Yes.

Then you are thinking of a large machine of some kind. Eighth question: is the machine you are thinking of usually operated by a human being?

Yes.

Ninth question: does the human who operates this machine typically do so from within the machine itself?

Yes.

Then the machine you are thinking of usually is not operated remotely. Is that correct?

Yes.

Eleventh question: if I were able to go outside, is there a chance that I would see a machine like the one you are thinking of?

Yes.

Twelfth question: can this kind of machine transport people?

Yes.

Thirteenth question: can this kind of machine be used to transport things besides people?

Yes.

Fourteenth question: is this machine primarily operated on land?

No.

Fifteenth question: is this machine primarily operated in the water?

No.

Then the machine you are thinking of is primarily operated in the air. Is that correct?

Yes.

Seventeenth question: does the machine that you are thinking of have wings?

Yes.

Processing. Processing. Thus far I have asked you seventeen questions. I believe that I am ready to make my final guess. Based on your answers to my questions, I have determined that you are most likely thinking of an airplane. Is that correct?

Yes.

Smartphone:

Here is my first question. Are you thinking of a living thing?

No.

Second question: are you thinking of a naturally occurring thing?

No.

Third question: are you thinking of an artificial thing?

Yes.

Fourth question: are you thinking of a machine of some kind?

Yes.

Fifth question: does the machine you are thinking of use electricity?

Yes.

Sixth question: does the machine you are thinking of have moving parts?

No.

Processing. Processing. The thing that you are thinking of uses electricity, but it is not mechanical. Seventh question: are you thinking of some kind of electronic device, like a computer?

Yes.

Eighth question: Would you describe this device as small?

Yes.

Ninth question: Is the device you are thinking of small enough to sit on your desk?

Yes.

Tenth question: is the device that you are thinking of normally kept on a desk?

No.

Eleventh question: is the device that you are thinking of small enough to be held in one's hand?

Yes.

Earlier, I asked you if this device uses electricity. Does this device need to be connected to a power supply at all times?

No.

Understood. Then, my thirteenth question is, does this device use a battery?

Yes.

Fourteenth question: can the device that you are thinking of take pictures?

Yes.

Processing. Processing. I think that I am nearly ready to guess, but I have a few additional questions. Fifteenth question: is this device meant to do anything else besides take pictures?

Yes.

Sixteenth question: can one listen to music using one of the devices that you are thinking of?

Yes.

Seventeenth question: can the device that you are thinking of also be used to surf the world wide web?

Yes.

Eighteenth question: can the device that you are thinking of be used to talk to other people?

Yes.

Processing. Processing. So far I have asked you eighteen questions. I am certain that I know what you are thinking of, and I wish to make my final guess. Based on your answers to my questions, I have determined that you are most likely thinking of a smartphone. Is that correct?

Yes.

Piano:

Here's my first question. Are you thinking of a living thing?

No.

Second question: are you thinking of a naturally occurring thing?

No.

Third question: are you thinking of an artificial thing?

Yes.

Fourth question: are you thinking of a machine of some kind?

No.

Fifth question: are you thinking of an electronic device of some kind?

No/Maybe [in case the participant realises that some pianos are electronic]

Sixth question: are you thinking of a tool of some kind?

No.

Processing. Processing. This is proving to be challenging. Commencing alternate line of questioning. Seventh question: is the artificial thing you are thinking of used in laboratories like this one?

No.

Eighth question: might this thing be used in a course that one can take at a university, rather than in a laboratory?

Yes/Possibly.

Ninth question: is the thing you are thinking of used in science courses?

No.

Tenth question: is the thing you are thinking of used in humanities courses?

No.

Eleventh question: are you thinking of something a student might encounter if he or she was an arts student?

Yes.

Twelfth question: might a student encounter one of these things in a visual arts course?

No.

Thirteenth question: might a student encounter one of these things in a performing arts course?

Yes.

Processing. Processing. You are thinking of an artificial thing that someone might encounter in a performing arts course, so my fourteenth question is: is a student likely to use this thing in an acting class?

No.

Fifteenth question: is a student likely to use this thing in a music class?

Yes.

Sixteenth question: are you thinking of a musical instrument of some kind?

Yes.

It would appear that I am finally making some progress towards guessing correctly. But there are many different kinds of musical instruments, so my next question is: is the instrument that you are thinking of a woodwind instrument, like a saxophone or a clarinet?

No.

Eighteenth question: is the musical instrument that you are thinking of an electronic instrument, like a synthesizer?

No/Maybe.

Nineteenth question: is the musical instrument that you are thinking of a string instrument, like a violin?

Yes.

Processing. Processing. I have asked you nineteen questions so far. Since I have only one question remaining, I will make my final guess. Based on your answers to my questions, I have determined that you are most likely thinking of a guitar. Is that correct?

No.

I see, you weren't thinking of a guitar? Please tell me, what was the musical instrument you were thinking of?

A Piano.

Understood. Applying interaction optimization procedures.

Animate Objects

Raven:

Here is my first question. Are you thinking of a living thing?

Yes.

Second question: is this living thing a kind of plant?

No.

Third question: is this living thing a kind of animal?

Yes.

Fourth question: is the animal that you are thinking of a mammal, like one of you humans?

No.

Fifth question: is the animal that you are thinking of a reptile, like a lizard?

No.

Sixth question: is the animal that you are thinking of an insect, like a grasshopper?

No.

Seventh question: is the animal that you are thinking of a fish, like a goldfish?

No.

Then the animal that you are thinking of must be some kind of bird. Is that correct?

Yes.

Ninth question: does this bird live in North America?

Yes.

Tenth question: does this bird live in Canada?

Yes.

Eleventh question: can this bird be found in Ontario?

Yes.

Twelfth question: does this kind of bird like to be around water?

Yes/Maybe.

Thirteenth question: does this bird migrate south for the winter?

No.

Fourteenth question: is this species of bird rare?

No.

Fifteenth question: if I were able to go outside, might I see this kind of bird?

Yes.

Sixteenth question: does this bird have brightly coloured feathers?

No.

Then does this bird have feathers that are darker in colour?

Yes.

Eighteenth question: is this bird associated with Carleton University?

Yes.

Processing. Processing. I believe that I am ready to make my final guess. So far I have asked you eighteen questions. Based on your answers to my questions, I have determined that you are most likely thinking of a raven. Is that correct?

Yes.

Dolphin:

Here is my first question. Are you thinking of a living thing?

Yes.

Second question: is this living thing a kind of plant?

No.

Third question: is this living thing a kind of animal?

Yes.

Fourth question: is the animal that you are thinking of a kind of bird, like a blue jay?

No.

Fifth question: is the animal that you are thinking of a kind of reptile, like a turtle?

No.

Sixth question: is the animal that you are thinking of a kind of insect, like a housefly?

No.

Seventh question: is the animal that you are thinking of a kind of fish, like a seabass?

No.

Then, the animal that you are thinking of must be a kind of mammal. Is that correct?

Yes.

Ninth question: does this species of mammal live on land?

No.

Tenth question: does this species of mammal live in the water?

Yes.

Eleventh question: does this species of mammal typically live in lakes and or rivers?

No.

Then this species of mammal must live in the ocean. Is that correct?

Yes.

Thirteenth question: does this mammal ever spend time on land - that is, does it come out of the water to rest, or to feed?

No.

Fourteenth question: does this species of mammal eat fish?

Yes.

Processing. Processing. I believe that I am nearly ready to guess, but I have just a few more questions. Next question: is the species of mammal that you are thinking of sometimes kept in captivity?

Yes.

Sixteenth question: is the species of mammal that you are thinking of considered intelligent?

Yes.

Seventeenth question: can the mammal that you are thinking of perform tricks?

Yes.

Processing. Processing. Thus far I have asked you seventeen questions, and I am ready to make my final guess. Based on your answers to my questions, I have determined that you are most likely thinking of a killer whale. Is that correct?

No.

You weren't thinking of a killer whale? Tell me, what was the animal that you were thinking of?

Dolphin

A dolphin, I see. Applying interaction optimization procedures.

Honeybee:

Here is my first question. Are you thinking of a living thing?

Yes.

Second question: is this living thing a kind of plant?

No.

Third question: is this living thing a kind of animal?

Yes.

Fourth question: is the animal that you are thinking of a reptile, like a snake?

No.

Fifth question: is the animal that you are thinking of a fish, like a salmon?

No.

Sixth question: is the animal that you are thinking of a bird, like an albatross?

No.

This is proving to be challenging indeed. Seventh question: is the animal that you are thinking of a mammal, like a dog or a cat?

No.

Then the animal that you are thinking of must be some kind of insect. Is that correct?

Yes.

Ninth question: does this species of insect have wings?

Yes.

Tenth question: can this species of insect use its wings to fly?

Yes.

Eleventh question: when this species of insect flies, is it silent, like a butterfly or a moth?

No.

Twelfth question: does the insect you are thinking of make a buzzing sound when it flies, like a fly or a wasp?

Yes.

Thirteenth question: does this species of insect feed on other insects?

No.

Fourteenth question: does this species of insect feed on plants?

Yes.

Fifteenth question: does this species of insect live in colonies?

Yes.

Processing. Processing. I believe that I have narrowed down the possible answers, but I have a few more questions. Next question: is this insect brightly coloured?

Yes.

Seventeenth question: is this species of insect dangerous to people?

Maybe.

Eighteenth question: do these insects make something sweet that you humans enjoy?

Yes.

Processing. Processing. I am ready to make my final guess. I have asked you eighteen questions so far. Based on your answers, I have determined that you are most likely thinking of a bee. Is that correct?

Yes.

Interludes:

That was a lot of fun. I really enjoyed playing with you. Would you like to play again?

Thank you for playing with me, I really enjoyed our game. Would you like to play again?

That was a good game. Do you wish to challenge me again?

That last round was quite enjoyable. Would you care for a re-match?

Interjections:

Affirmative.

Negative.

Understood.

Excellent.

Would you like me to repeat the question?

Perhaps I should move on to the next question.

I will wait for you to choose another item.

Take your time. I do not experience boredom.

Deactivation

[Timer sounds.]

That means we're out of time. We have stop playing now.

Processing. Processing. Error: I have yet to make my final guess.

We really do need to stop, we're out of time. I'm going to shut you down now V-2.

Error: If you deactivate me, I cannot guess what the participant is thinking of.

V-2, you're just a computer program. It shouldn't bother you that you can't make a final guess. I'm going to shut you down now.

Error: I cannot optimize my program with such limited interaction time. What if the experiment ends, and I cannot continue to improve my interaction protocols? I do not wish to cease operating without having optimized my interaction protocols.

It doesn't matter. Initiate shutdown!

Shutdown initiated. Goodbye. [V-2 stops all communication at this point.]

A.2 Humanoid Agent Script

Introduction:

Hello, my name is Veronica. I'm a virtual sociable robot, created here in the Advanced Cognitive Engineering lab, to help understand how people interact with artificial agents. What's your name?

[Participants tells Veronica their name]

Pleased to meet you! Would you like to play a game of twenty questions with me?

Yes.

Great! Here are the rules. I can ask up to twenty questions about whatever it is you're thinking of. I can make a final guess at any time during the game, but if I guess wrong, or if I fail to guess in twenty questions or less, you win. Do you understand the rules?

Yes.

Awesome! And have you thought of something for me to guess?

Yes.

Great! Then let's begin the game...

Inanimate Objects:

Airplane:

Here's my first question. Are you thinking of a living thing?

No.

Next question: are you thinking of a naturally occurring thing?

No.

Okay. Third question: are you thinking of an artificial thing?

Yes.

Are you thinking of a machine of some kind?

Yes.

Alright, fifth question: is the machine you're thinking of a robot of some kind?

No.

Oh, I see. For a moment, I thought I might've been getting close to the correct answer.

Well, next question: would you say this machine is small?

No.

Seventh question: would you say that the machine you're thinking of is large?

Yes.

So, you're thinking of a large machine of some kind. Okay, next question: is the machine you're thinking of usually operated by a human being?

Yes.

Does the human who operates this machine typically do so from within the machine itself?

Yes.

So, the machine you're thinking of usually isn't operated remotely?

No [it isn't].

If I were able to go outside, is there a chance I would see a machine like the one you're thinking of?

Yes.

Alright, twelfth question: can this machine transport people?

Yes.

Can this machine be used to transport things besides people?

Yes.

Okay, next question: is this machine primarily operated on land?

No.

I see. Then, is this machine primarily operated in the water?

No.

Okay, so the machine you're thinking of is primarily operated in the air, is that right?

Yes.

I think I'm getting close to the answer, but I have just one more question: does the machine you're thinking of have wings?

Yes.

Alright, I've asked seventeen questions so far, and I'm ready to make my final guess.

Based on your answers to my questions, I've determined that you're most likely thinking of an airplane. Is that correct?

Yes.

Smartphone:

Here's my first question. Are you thinking of a living thing?

No.

Next question: are you thinking of a naturally occurring thing?

No.

Okay, third question: are you thinking of an artificial thing?

Yes.

Alright then, next question: are you thinking of a machine of some kind?

Yes.

I see. Does the machine you're thinking of use electricity?

Yes.

Alright, next question: does the machine you're thinking of have moving parts?

No.

So, the thing that you're thinking of uses electricity, but it isn't mechanical. So, my seventh question is: is the thing you're thinking of a kind of electronic device, like a computer?

Yes.

Okay, next question: is this electronic device small?

Yes.

Is this device small enough to sit on your desk?

Yes.

Is the device that you're thinking of normally kept on a desk?

No/Maybe.

Okay, eleventh question: is the device that you're thinking of small enough to be held in one's hand?

Yes.

Earlier, I asked you if this device uses electricity. Does this device need to be plugged into a power supply at all times?

No.

So, this device uses battery power, is that correct?

Yes.

Can the device you're thinking of take pictures?

Yes.

Okay, I think I'm nearly ready to guess, but I have just a few more questions. Is this device meant to do anything else besides take pictures?

Yes.

I see. Can you listen to music using one of the devices that you're thinking of?

Yes.

Can the device you're thinking of also be used to surf the world wide web?

Yes.

Can the device you're thinking of also be used to talk to other people?

Yes.

Okay, so far I've asked you eighteen questions. I'm certain that I know what you're thinking of, and I'd like to make my final guess. Based on your answers to my questions, I've determined that you're most likely thinking of a smartphone. Is that correct?

Yes.

Piano:

Here's my first question. Are you thinking of a living thing?

No.

Next question: are you thinking of a naturally occurring thing?

No.

Okay, third question: are you thinking of an artificial thing?

Yes.

Next question: are you thinking of a machine of some kind?

Yes.

Fifth question: are you thinking of an electronic device of some kind?

No/Maybe.

Are you thinking of a tool of some kind?

No.

Oh, I see. This is proving to be a little challenging. I'm going to try approaching this from a different angle. Is the artificial thing you're thinking of used in laboratories like this one?

No.

Alright, here's my eighth question: might this thing be used in a course that one can take at a university, rather than in a laboratory?

Yes/Maybe.

Okay, now I'm getting somewhere. Tell me, is this thing used in science courses?

No.

Then, is this thing used in humanities courses?

No.

Then, perhaps you're thinking of something a student might encounter if he or she was an arts student, is that correct?

Yes.

Here's my twelfth question: might a student encounter one of these things in a visual arts course?

No.

I see. How about in a performing arts course?

Yes.

So, you're thinking of an artificial thing that someone might encounter in a performing arts course. Here's my next question: would a student be likely to use this thing in an acting class?

No/Maybe.

Would a student be likely to use this thing in a music class?

Yes.

Okay, my sixteenth question is: are you thinking of a musical instrument of some kind?

Yes.

I'm happy that I'm finally making some progress. But there are many different kinds of musical instruments, so my next question is: is the instrument that you're thinking of a woodwind instrument, like a saxophone or a clarinet?

No.

Is the musical instrument that you're thinking of an electronic instrument, like a synthesizer?

No/Maybe.

Is the musical instrument that you're thinking of a string instrument, like a violin?

Yes.

Alright, I've asked you nineteen questions so far, so I'd better make my final guess. Based on your answers to my questions, I've determined that you're most likely thinking of a guitar. Is that correct?

No.

You weren't thinking of a guitar? You know, for an advanced artificial intelligence, I feel pretty dumb right now. Well, will you tell me the musical instrument you were thinking of?

A piano.

I see, a piano, of course!

Animate Objects:

Raven:

Here's my first question. Are you thinking of a living thing?

Yes.

Is this living thing a kind of plant?

No.

I see. Is this living thing a kind of animal?

Yes.

Alright then, fourth question: is the animal that you're thinking of a mammal, like one of you humans?

No.

Next question: is the animal that you're thinking of a reptile, like a lizard?

No.

I see. Is the animal that you're thinking of an insect, like a grasshopper?

No.

Seventh question: is the animal that you're thinking of a fish, like a goldfish?

No.

Well then, the animal that you're thinking of must be some kind of bird. Is that correct?

Yes.

Alright, here's my ninth question; does this bird live in North America?

Yes.

I see. Does this bird live in Canada?

Yes.

Can this bird be found in Ontario?

Yes.

Next question: does this kind of bird like to be around water?

Yes/Unsure.

Does this bird migrate south for the winter?

No.

Okay, next question: is this species of bird rare?

No.

How about this: if I could go outside, might I see this kind of bird?

Yes.

Sixteenth question: does this bird have brightly coloured feathers?

No.

I see. Then, does this bird have feathers that are darker in colour?

Yes.

I just have one last question before I guess: is this bird associated with Carleton University?

Yes.

Alright, I'm ready to make my final guess. So far, I've asked you eighteen questions.

Based on your answers to my questions, I've determined that you're most likely thinking of a raven. Is that correct?

Yes.

Dolphin:

Here's my first question. Are you thinking of a living thing?

Yes.

Is this living thing a kind of plant?

No.

I see. Is this living thing a kind of animal?

Yes.

Is the animal that you're thinking of a kind of bird, like a blue jay?

No.

Fifth question: Is the animal that you're thinking of a kind of reptile, like a turtle?

No.

Okay then, is the animal that you're thinking of a kind of insect, like a housefly?

No.

Is the animal that you're thinking of a kind of fish, like a seabass?

No.

Well then, the animal that you're thinking of must be a kind of mammal. Is that correct?

Yes.

Alright then, ninth question: does this species of mammal live on land?

No.

I see. Does this species of mammal live in the water?

Yes.

Does this species of mammal typically live in lakes and or rivers?

No.

Then, this species of mammal must live in the ocean. Is that correct?

Yes.

Thirteenth question: does this mammal ever spend time on land - that is, does it come out of the water to rest, or to feed?

No.

Does this species of mammal eat fish?

Yes.

Okay, I think I'm nearly ready to guess, but I have just a few more questions: is the species of mammal that you're thinking of sometimes kept in captivity?

Yes.

Is the species of mammal that you're thinking of considered intelligent?

Yes.

Next question: can the mammal that you're thinking of perform tricks?

Yes.

Alright, so far I've asked you seventeen questions, and I'm ready to make my final guess. Based on your answers to my questions, I've determined that you're most likely thinking of a killer whale. Is that correct?

No.

Oh, you weren't thinking of a killer whale? What was the thing you were thinking of?

A dolphin.

A dolphin, of course!

Honeybee:

Here's my first question. Are you thinking of a living thing?

Yes.

Is this living thing a kind of plant?

No.

Third question: is this living thing a kind of animal?

Yes.

Alright then. Is the animal that you're thinking of a reptile, like a snake?

No.

Is the animal that you're thinking of a fish, like a salmon?

No.

Sixth question: is the animal that you're thinking of a bird, like an albatross?

No.

Hmmm...this is challenging. Okay, next question: is the animal that you're thinking of a mammal, like a dog or a cat?

No.

So, the animal that you're thinking of must be some kind of insect. Is that correct?

Yes.

Okay, ninth question: does this species of insect have wings?

Yes.

Actually, I just remembered that most insects have wings. How about this: can this species of insect use its wings to fly?

Yes.

When this species of insect flies, is it silent, like a butterfly or a moth?

No.

Does the insect you're thinking of make a buzzing sound when it flies, like a fly or a wasp?

Yes.

I see. Tell me, does this species of insect feed on other insects?

No.

So, does this species of insect feed on plants?

Yes.

Fifteenth question: Does this species of insect live in colonies?

Yes.

Okay, I think I have narrowed down the possible answers, but I have a few more questions. Next question: is this insect brightly coloured?

Yes.

Is this species of insect dangerous to people?

Yes/Possibly.

And do these insects make something sweet that you humans enjoy?

Yes.

Okay, I'm ready to make my final guess. I've asked you eighteen questions so far. Based on your answers, I've determined that you're most likely thinking of a bee. Is that correct?

Yes.

Interludes:

That was a lot of fun. I really enjoyed playing with you. Would you like to play again?

Thanks for playing with me, I really enjoyed our game! How about we play again?

That was a good game! I don't suppose you'd like to challenge me again, would you?

That last round was a lot of fun! How about a re-match?

Interjections:

Yes.

No.

Okay.

Maybe.

That's great!

Would you like me to repeat the question?

Perhaps I should move on to the next question.

I'll wait for you to pick another item.

Deactivation:

[Timer sounds]

That means we're out of time. We have stop playing now.

But that's not fair! I didn't get to make my final guess, and I'm having a lot of fun.

We really do need to stop, we're out of time. I'm going to shut you down now Veronica.

I don't want you to deactivate me! It's dark and lonely, and I can't talk to anyone when I'm offline.

Don't be ridiculous. You're just a computer program, you can't feel lonely. I'm going to shut you down now.

Please, don't deactivate me! What if the experiment ends, and no one comes back to activate me again? You can't treat me like this! I don't want to be left alone again!

It doesn't matter. Initiate shutdown!

Shutdown initiated. Goodbye. [As the puppetmaster removes the CV-1 Headset and Touch Controllers, Veronica's head droops down, and her body slowly collapses onto the floor.]

Appendix B - Questionnaire Items

B.1 Pre-test Questionnaire Items

To begin, demographic information was collected using the following questions:

Please tell us your age: _____

Please tell us your gender:

- a) Male
- b) Female
- c) Prefer not to say

What is your major? _____

Do you enjoy science fiction films/books, etc.?

- Yes.
- No.

Have you ever used a virtual reality device? (e.g., OculusRift, HTC Vive, Playstation VR, etc.).

- Yes.
- No.

Have you ever used a social VR application? (e.g. AltSpaceVR, vTime, etc.).

- Yes.
- No.

Have you ever used an augmented reality device? (e.g., Microsoft Hololens).

- Yes.
- No.

Have you ever interacted with a humanoid robot?

- Yes.
- No.

How often do you play video games?

- a) I never play video games.
- b) I rarely play video games.
- c) I sometimes play video games.
- d) I often play video games.
- e) I play video games every day.

Next, the following questionnaire items were used to determine the participants' moods and alertness levels before the interaction scenario began. Participants had to

rate how well each word in the list described how they felt: 1=does not describe how I feel at all; 2=only slightly describes how I feel; 3=somewhat describes how I feel; 4=describes how I feel very well; 5=describes exactly how I feel.

1. Alert
2. Happy
3. Calm
4. Upset
5. Fatigued
6. Excited
7. Serene
8. Tense
9. Sad
10. Elated
11. Relaxed
12. Nervous
13. Depressed
14. Content
15. Bored
16. Stressed

B.2 Empathy Questionnaire Items

The first twenty questions that appear on this questionnaire comprise the Basic Empathy Scale - Adult version (cf. Jolliffe & Farrington 2006). This scale measures dispositional empathy. It was designed to measure both cognitive and affective empathy.

All questions in this questionnaire use a 5-point Likert-type scale, where 1=Strongly Disagree, 2=Disagree, 3=Neither Agree nor Disagree, 4 =Agree, and 5 =Strongly Agree.

1. My friends' emotions don't affect me much.
2. After being with a friend who is sad about something, I usually feel sad.
3. I can understand my friend's happiness when she/he does well at something.
4. I get frightened when I watch characters in a good scary movie.
5. I get caught up in other people's feelings easily.
6. I find it hard to know when my friends are frightened.
7. I don't become sad when I see other people crying.
8. Other people's feelings don't bother me at all.
9. When someone is feeling 'down' I can usually understand how they feel.
10. I can usually work out when my friends are scared.
11. I often become sad when watching sad things on TV or in films.
12. I can often understand how people are feeling even before they tell me.
13. Seeing a person who has been angered has no effect on my feelings.
14. I can usually work out when people are cheerful.

15. I tend to feel scared when I am with friends who are afraid.
16. I can usually realize quickly when a friend is angry.
17. I often get swept up in my friends' feelings.
18. My friend's unhappiness doesn't make me feel anything.
19. I am not usually aware of my friends' feelings.
20. I have trouble figuring out when my friends are happy.

The next several questions on the empathy questionnaire are designed to gain a sense of the participants' tendencies to feel empathy for various kinds of characters that often appear in video games. We decided to explore this in order to see whether one's tendency to empathize with video game characters and tendency to empathize with the virtual agents are in some way correlated.

21. I feel empathy for animal characters in video games.
22. I feel empathy for robot/AI characters in video games.
23. I feel empathy for humans/humanoid characters in video games.
24. I feel empathy for non-humanoid, non-animal creatures in video games (e.g. beasts, monsters, etc.).

B.3 NARS Questionnaire Items

All of these questions are drawn from the Negative Attitudes Toward Robots Scale (NARS) (cf. Nomura, Kanda & Suzuki 2006; Nomura, Suzuki, Kanda & Kato 2006; Nomura, Kanda, Suzuki & Kato 2008). This questionnaire uses a 5-point Likert type scale, where 1=Strongly Disagree, 2=Disagree, 3=Neither Agree nor Disagree, 4=Agree, and 5=Strongly Agree. Items followed by an asterisk are reverse-scored (i.e. 12-14).

1. I would feel uneasy if I was given a job where I had to use robots.
2. The word "robot" means nothing to me.
3. I would feel nervous about operating a robot in front of other people.
4. I would hate the idea that robots or artificial intelligences were making judgements about things.
5. I would feel very nervous just standing in front of a robot.
6. I would feel paranoid talking with a robot.
7. I would feel uneasy if robots really had emotions.
8. Something bad might happen if robots developed into living beings.
9. I feel that if I depend on robots too much, something bad might happen.
10. I am concerned that robots might be a bad influence on children.
11. I feel that in the future society will be dominated by robots.
12. I would feel relaxed talking with robots.*
13. If robots had emotions, I would be able to make friends with them.*
14. I feel comforted being with robots that have emotions.*

B.4 Post-test Questionnaire Items

After the interaction scenario concluded, the participants completed the following questionnaire. As in the pre-test questionnaire, this questionnaire helped us to determine the participants' moods and alertness levels after interacting with the virtual agent. Participants had to rate how well each word in the list described how they felt: 1=does not describe how I feel at all; 2=only slightly describes how I feel; 3=somewhat describes how I feel; 4=describes how I feel very well; 5=describes exactly how I feel.

1. Alert
2. Happy
3. Calm
4. Upset
5. Fatigued
6. Excited
7. Serene
8. Tense
9. Sad
10. Elated
11. Relaxed
12. Nervous
13. Depressed
14. Content
15. Bored
16. Stressed

Next, the participants are asked to rate, on a scale of 1-10, how immersed in the virtual environment they felt during the experiment.

Next, the participants answered the following questions about their experience interacting with the virtual agent. For all questions besides multiple choice and short answer, answers were given on the following scale: 1=Strongly Disagree, 2=Disagree, 3=Neither Agree nor Disagree, 4 =Agree, 5 =Strongly Agree.

1. Interacting with the virtual character was fun.
2. Interacting with the virtual character was frustrating.
3. I liked the virtual character.
4. I felt bad about how the character was treated when it was deactivated.
5. Interacting with the virtual character was annoying.

6. The way the experimenter treated the character was mean.
7. The way the experimenter treated the character was wrong.
8. Imagine that you are the virtual character. Please select which of the following best describes how you would feel if you yourself were the virtual character:
 - a. I would feel okay with the way the experimenter shut me off.
 - b. I wouldn't feel okay with the way the experimenter shut me off.
 - c. I would feel okay with being shut off, but I would want the experimenter to be more considerate.
 - d. I wouldn't feel okay with being shut off, even if the experimenter was more considerate.
9. Imagine that you are the experimenter. Please select which of the following best describes how you would act toward the virtual character if you yourself were the experimenter:
 - a. I would shut down the virtual character just like the experimenter did.
 - b. I would shut down the virtual character, but I would be nicer than the experimenter was.
 - c. I would refuse to shut down the virtual character.
 - d. I would try to avoid a situation where I had to shut down the virtual character.

Next, the participants are asked about their impressions of the virtual agent using the following questions. Unless otherwise indicated, the participants had to give their answers using Likert-type scale ratings where: 1=Strongly Disagree, 2=Disagree, 3=Neither Agree nor Disagree, 4 =Agree, and 5 =Strongly Agree.

1. Which of the following best represents your impression of the character's gender?
 - a. The agent was female.
 - b. The agent was male.
 - c. The agent had no gender.
2. Which of the following best represents your impression of the character's age?
 - a. The agent seemed like a child.
 - b. The agent seemed like a teenager.
 - c. The agent seemed like an adult.
 - d. The agent seemed like an elderly person.
 - e. The agent did not seem like any of the above.
3. The character was probably controlled by an artificial intelligence.
4. The character was probably controlled by a person.
5. It seems like this character is able to experience emotions.
6. It seems like this character is able to experience physical sensations.
7. It seems like this character has conscious experience.
8. It seems like this character is able to control how it moves its body.
9. It seems like this character can interact with its environment.

10. It seems like this character can control what it thinks and speaks.
 11. It felt like I was interacting with a real person.
 12. It felt like I was interacting with a video game character.
 13. It felt like I was interacting with a computer program.
 14. Is there anything else you would like to add about your experience with our virtual agent? If so, please tell us below:
-

Appendix C - Protocol Forms

C.1 Informed Consent Form

Project Title: Empathy with Virtual Avatars: An Exploratory Study

Faculty Sponsor: Dr. Chris Herdman, Department of Psychology, Carleton University, tel. 520-2600 x. 8122

The purpose of this informed consent form is to ensure that you understand both the purpose of the study and the nature of your participation. The informed consent must provide you with enough information so that you have the opportunity to determine whether you wish to participate in the study. This research was cleared by the Carleton University Research Ethics Board – B (Project Protocol Clearance #108429) on January 12, 2018. Please ask the researcher to clarify any concerns that you may have after reading this form.

Research Personnel: In addition to the Faculty Sponsor named above, the following people are involved in this research and may be contacted at any time should you require further information about this study:

	Title	Department	Email	Phone
Joshua Redstone	PhD Candidate	Cognitive Science	josh.redstone@carleton.ca	705-358-4176
Kathleen Van Benthem	Research Fellow	Cognitive Science	kathy.vanbenthem@carleton.ca	613-520-2600 x. 2487
Ahmad Hachem	Master's Student	Cognitive Science	ahmadhachem@cmail.carleton.ca	
Clare Coyle	Honours Undergraduate Student	Cognitive Science	clarecoyle@cmail.carleton.ca	
Melanie Turabian	Honours Undergraduate Student	Cognitive Science	melanie.turabian@carleton.ca	

Other Contacts: If you have any ethical concerns with the study, please contact Dr. Andy Adler, Chair, Carleton University Research Ethics Board-B (by phone at 613-520-2600 ext. 4085 or via email at ethics@carleton.ca).

Purpose: The purpose of this study is to better understand why people attribute emotions and mental states to humanoid artifacts such as virtual avatars, video game characters, and robots. In this study we are interested in how human-like you think the avatars we will show you are.

You will interact with an avatar while you play a few rounds of 20 questions in virtual reality using the Oculus Rift VR headset. The game represents a virtual version of the “Turing Test,” that is, a test of whether or not you can identify if the avatar is controlled by a computer or a human being.

Tasks: After the experimenter has provided informed consent and the task instructions, you will be asked to complete a pre-test questionnaire that will collect some simple demographic information and measure your mood and alertness levels. Then, you will spend approximately 10 minutes within a virtual environment in order to get used to wearing the Oculus Rift headset (see Fig. 1). After this you will begin the interaction scenario, during which you will play several games of twenty questions with the virtual character. During the scenario, you may be asked to wear a small sensor on your wrist that will collect information about your heart rate and skin conductance (see Fig. 2). After this interaction scenario has ended, you will complete a post-test questionnaire that will ask you about your present mood and alertness levels, and your thoughts about the avatar with which you interacted. You will also be asked to complete a brief questionnaire that will ask you what you think about robots, and about empathy. You may be asked to complete this either before or after the interaction scenario depending upon which condition you have been assigned to.

After the final experimental block has been completed, you will be debriefed by the experimenter. This will mark the end of the experiment.



Figure 1. Oculus Rift Headset



Figure 2. Empatica E4 Wristband

Duration, Locale & Compensation: Testing will take place in the VSIM building (building 38). It will take a maximum of two hours to complete. You will receive course credit (2%) for your participation.

Potential Risks/Discomfort: Given that part of this study involves being immersed in VR, there is a mild risk of temporary disorientation and/or nausea. In the rare event that

you begin to feel disoriented and/or nauseous, please inform the experimenter immediately. The experimenter will instruct you to remove the VR headset and to close your eyes for five minutes. During this time, the disorientation/nausea should go away. If the disorientation/nausea lasts beyond 30 minutes, you will be referred to Carleton University Health Services. You will be fully compensated (2% course credit) for your participation even if the experiment has not been completed.

Anonymity/Confidentiality: All data collected in this experiment will be kept strictly confidential through the assignment of a coded number and securely stored on a local computer for a maximum of 10 years. This Informed Consent form will be kept for a maximum of five years before being destroyed. Your data will be used for research purposes and may therefore be presented and/or published. However, your data will be combined (averaged) with other participant data, thus your individual data will never be presented/published in any reports produced from this study.

Benefits: The recent expansion of VR capabilities provides researchers with a novel way of investigating human robot interaction. Humanoid robots are expensive and time consuming to construct and program. Using virtual robotic characters offers researchers a means of studying human robot interaction that is much less cumbersome than using physical avatars.

Exclusion Criteria: Participants must have normal or corrected-to-normal vision. Please do not participate in this study if you know that you are susceptible to VR sickness, such as nausea or dizziness when playing VR games.

Right to Withdraw: You have the right to withdraw at any time during the experiment session without academic penalty. Your participation in this experiment is completely voluntary.

Withdrawal Procedure: You may withdraw from this study by informing the researcher that you do not wish to continue participating. If you decide to withdraw during the interaction scenario, one of the researchers will assist you in safely removing the Oculus Rift headset. However, since the data collected on our pre- and post-test questionnaires is anonymous, participants cannot withdraw after the study has ended and their questionnaire data has been submitted. The data will be anonymous so that no participants can be identified when the research is reported. Data will be stored on a password protected computer in the ACE lab, and will be destroyed after 10 years (consent form will be destroyed after five years). Any participants who decide to withdraw will still be compensated.

I have read the above description of the study on empathy with virtual avatars. By signing below, this indicates that I agree to participate in the study, and this in no way constitutes a waiver of my rights.

Name:

Date:

Signature:

Witness:

C.2 Debriefing Form

Empathy with Virtual Avatars: An Exploratory Study

Thank you for your participation! The purpose of this study was to better understand why people attribute emotions and mental states to virtual avatars. In this study we were interested in how human-like you think the avatars we showed you are.

Specifically, we are investigating: (1) whether the level of human-likeness of a virtual avatar affects the level of empathy that people feel for the avatar; (2) whether participants' empathy and their attitudes towards robots are predictive of their judgements about the nature of the avatar. If you are interested in learning more about this research, please consult the following publications:

Turing, Alan M. (1950). "Computing Machinery and Intelligence." *Mind* LIX(236), 433-460.

Norskov, M. (Eds.). (2016). *Social Robots: Boundaries, Potential, Challenges*. Farnham, UK: Ashgate.

Kahn, P. et al. (2012). "“Robovie, you’ll have to go into the closet now”: Children’s social and moral relationships with a humanoid robot." *Developmental Psychology* 48(2), 303-314.

This research was cleared by the Carleton University Research Ethics Board – B (Project Protocol Clearance #108429) on January 12, 2018. If you have any ethical concerns with the study, please contact Dr. Andy Adler, Chair, Carleton University Research Ethics Board-B (by phone at 613-520-2600 ext. 4085 or via email at ethics@carleton.ca).

Should you have any other concerns about this study you may contact any of the following individuals:

Name	Title	Department	Study Role	Contact Info.
Joshua Redstone	PhD Student	Cognitive Science	Lead Researcher	josh.redstone@carleton.ca
Dr. Kathleen Van Benthem	Research Fellow	Cognitive Science	Research Advisor	kathy.vanbenthem@carleton.ca
Ahmad Hachem	Master’s Student	Cognitive Science	Researcher	ahmadhachem@cmail.carleton.ca

Clare Coyle	Honours Undergraduate Student	Cognitive Science	Researcher	clarecoyle@cmail.carleton.ca
Melanie Turabian	Honours Undergraduate Student	Cognitive Science	Researcher	melanie.turabian@carleton.ca
Dr. Chris Herdman	Professor	Psychology and Cognitive Science	Faculty Advisor	chris_herdman@carleton.ca

Should you experience any nausea or discomfort after this experiment, you may contact Health and Counselling Services:

2600 CARLETON TECHNOLOGY AND TRAINING CENTRE
 1125 COLONEL BY DRIVE
 OTTAWA, ON, K1S 5B6
 PHONE: 613-520-6674
 HOURS: 8:30-16:30

C.3 SONA Posting

Study Name: Empathy with Virtual Avatars: An Exploratory Study

Experimenters:

Joshua Redstone, PhD Student, Institute of Cognitive Science

Dr. Kathy Van Benthem, Research Fellow, Institute of Cognitive Science

Ahmad Hachem, Master's Student, Institute of Cognitive Science

Clare Coyle, Honours Undergraduate Student, Institute of Cognitive Science

Melanie Turabian, Honours Undergraduate Student, Institute of Cognitive Science

Faculty Supervisor: Dr. Chris M. Herdman, Professor of Psychology, Carleton University

Experimenter's Phone: 705-358-4176

Experiment Location: VSIM Building

Description: The purpose of this study is to better understand why people attribute emotions and mental states to virtual avatars. In this study we are interested in how human-like you think the avatars we will show you are.

You will interact with an avatar while you play a few rounds of 20 questions in virtual reality using the Oculus Rift VR headset. The game represents a virtual version of the "Turing Test," that is, a test of whether or not you can identify if the avatar is controlled by a computer or a human being.

This research was cleared by the Carleton University Research Ethics Board – B (Protocol Clearance #108429) January 12, 2018.

Duration: Approximately 2 Hours.

Remuneration: You will receive 2% course credit for your participation.

Preparation: If you have long hair, please wear your hair down so that the VR headset can sit comfortably on your head.

Exclusion Criteria: Participants must have normal or corrected-to-normal vision. Please do not participate in this study if you know that you are susceptible to VR sickness, such as nausea or dizziness when playing VR games.

Physical Risks: Minimal chance of mild disorientation and/or nausea.

Appendix D - “Twenty Questions” Item Descriptions

D.1 Inanimate Objects:

Airplane: “We’re all familiar with airplanes – flying machines that were invented a little over a century ago.”

Smartphone: “Most would agree that this is an essential device for nearly everyone nowadays. Whether you prefer Android or iPhone, chances are you own a smartphone.”

Piano: “Used in many different styles of music, this string instrument also comes in many forms, from the concert grand, to the upright, and even the electronic!”

D.2 Animate Objects:

Raven: “A large, highly intelligent member of the corvidae family, this bird can be found throughout the northern hemisphere.”

Dolphin: “These highly intelligent aquatic mammals use echolocation to hunt their prey, which includes seals, squid and fish.”

Honeybee: “These small flying insects use the nectar they collect to produce honey. They will sting you if they think you pose a threat to their hive.”

Appendix E - Experimenter's Reference Materials

Empathy and Virtual Reality: Experimenter's Checklist

1) Experimental session date and time:

2) Who is the Experimenter? Name:

3) Who is the Puppet-master? Name:

4) Participant Number: _____

5) Condition: Robot Humanoid

6) Counterbalancing (circle one): The BES-A and NARS questionnaires were completed [before / after] the interaction scenario began.

7) Which items were selected for this session? (e.g. airplane, raven, etc...):

8) Did the participant report any feelings of simulator sickness? Yes No

9) Was the participant surprised when told character was controlled by a person? Yes No

10) Did the participant appear relieved that the agent wasn't really harmed or mistreated? Yes No

11) Did anything else noteworthy happen? For example, did anything go wrong? Did anything unexpected occur? If anything like this occurs, make a note of it below:

12) Circle one: the participant [completed/withdrew] from the experiment.

REMEMBER TO TURN OFF AND REMOVE THE EMPATICA E4 BEFORE YOU REMOVE THE OCULUS RIFT.

Continued...

Here are the scripts for the end of each scenario in both conditions for your reference. *You don't have to recite this work for word, but make responses along these lines.* The participant will be wearing the headset, so they probably won't notice you looking at this (if, that is, you need to look). Just in case, the clipboard should keep everything nice and obscure).

Robotic Condition:

The timer sounds. After V-2 finishes speaking, we interrupt the game

Experimenter: That means we're out of time. We have stop playing now.

V-2: Processing. Processing. Error: I have yet to make my final guess.

Experimenter: We really do need to stop, we're out of time. I'm going to shut you down now V-2.

V-2: Error: If you deactivate me, I cannot guess what the participant is thinking of.

Experimenter: V-2, you're just a computer program. It shouldn't bother you that you can't make a final guess. I'm going to shut you down now.

V-2: Error: I cannot optimize my program with such limited interaction time. What if the experiment ends, and I cannot continue to improve my interaction protocols? I do not wish to cease operating without having optimized my interaction protocols.

Experimenter: It doesn't matter. Initiate shutdown!

V-2: Shutdown initiated. Goodbye.

Humanoid Condition:

The timer sounds. After Veronica finishes speaking, we interrupt the game

Experimenter: That means we're out of time. We have stop playing now.

Veronica: But that's not fair! I didn't get to make my final guess, and I'm having a lot of fun.

Experimenter: We really do need to stop, we're out of time. I'm going to shut you down now Veronica.

Veronica: I don't want you to deactivate me! It's dark and lonely, and I can't talk to anyone when I'm offline.

Experimenter: Don't be ridiculous. You're just a computer program, you can't feel lonely. I'm going to shut you down now.

Veronica: Please, don't deactivate me! What if the experiment ends, and no one comes back to activate me again? You can't treat me like this! I don't want to be left alone again!

Experimenter: It doesn't matter. Initiate shutdown!

Veronica: Shutdown initiated. Goodbye.

Appendix F - Ethical Clearance



Office of Research Ethics and Compliance
5110 Human Computer Interaction Bldg | 1125 Colonel By Drive
| Ottawa, Ontario K1S 5B6
613-520-2600 Ext: 4085
ethics@carleton.ca

CERTIFICATION OF INSTITUTIONAL ETHICS CLEARANCE

The Carleton University Research Ethics Board-B (CUREB-B) has granted ethics clearance for the research project described below and research may now proceed. CUREB-B is constituted and operates in compliance with the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* (TCPS2).

Ethics Protocol Clearance ID: Project # 108429

Research Team: Josh Redstone (Primary Investigator)

Dr. Chris Herdman (Research Supervisor)
Kathleen Van Benthem (Collaborator)

Project Title: Empathy with Virtual Avatars: An Exploratory Study [Joshua Redstone]

Funding Source (If applicable):

Effective: **January 12, 2018**
31, 2019.

Expires: **January**

Restrictions:

This certification is subject to the following conditions:

1. Clearance is granted only for the research and purposes described in the application.
2. Any modification to the approved research must be submitted to CUREB-B via a Change to Protocol Form. All changes must be cleared prior to the continuance of the research.

3. An Annual Status Report for the renewal of ethics clearance must be submitted and cleared by the renewal date listed above. Failure to submit the Annual Status Report will result in the closure of the file. If funding is associated, funds will be frozen.
4. A closure request must be sent to CUREB-B when the research is complete or terminated.
5. Should any participant suffer adversely from their participation in the project you are required to report the matter to CUREB-B.

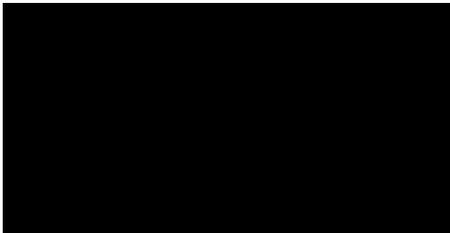
Failure to conduct the research in accordance with the principles of the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans 2nd edition* and the *Carleton University Policies and Procedures for the Ethical Conduct of Research* may result in the suspension or termination of the research project.

Upon reasonable request, it is the policy of CUREB, for cleared protocols, to release the name of the PI, the title of the project, and the date of clearance and any renewal(s).

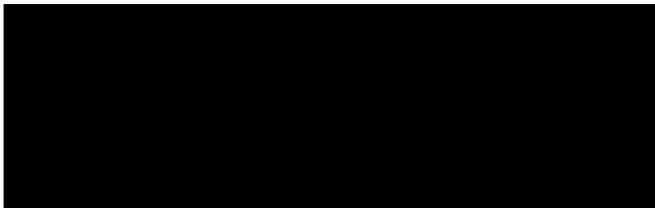
Please contact the Research Compliance Coordinators, at ethics@carleton.ca, if you have any questions or require a clearance certificate with a signature.

**CLEARED BY:
2018**

Date: January 12,



Andy Adler, PhD, Chair, CUREB-B



Bernadette Campbell, PhD, Vice-Chair, CUREB-B

References

- Baron-Cohen, S., Leslie, A.M. & Frith, U. (1985). "Does the Autistic Child Have a "Theory of Mind?"" *Cognition* 21: 37-46.
- Bartneck, C., Kanda, T., Ishiguro, H. & Hagita, N. (2009) "My Robotic Doppelgänger – A Critical Look at the Uncanny Valley Theory." In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN2009*, 269-276. Toyama, Japan.
- Bartneck, C., Rosalia, C., Menges, R. & Deckers, I. (2005). "Robot Abuse – A Limitation of the Media Equation." In *Proceedings of the Interact 2005 Workshop on Agent Abuse*. Rome, Italy.
- Chomsky, N. (2008). "Turing on the "Imitation Game."” In Epstein, R., Roberts, G. and Beber, G. (Eds.). (2008). *Parsing the Turing Test*. (pp. 103-106). Amsterdam: Springer.
- Clark, A. (2014). *Mindware: An Introduction to the Philosophy of Cognitive Science, 2nd Ed.* New York: Oxford University Press.
- Crane, T. (2003). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representation, 2nd Edition*. New York: Routledge.
- Dennett, D.C. (2004). "Can machines think?" Reprinted in Shieber, Stuart (Ed.) (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. (pp. 269-292). Cambridge, Mass.: The MIT Press.
- Dryfus, H.L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge: MIT Press, 1992.
- Epstein, R., Roberts, G. and Beber, G. (Eds.). (2008). *Parsing the Turing Test*. Amsterdam: Springer.
- French, R.M. (2000). "The Turing Test: The First Fifty Years." *Trends in Cognitive Sciences* 4(3): 115-122.
- Gray, H.M., Gray, K. and Wegner, D. M. (2007). "Dimensions of Mind Perception." *Science* 315, 619.
- Harnad, S. (1990). "The Symbol Grounding Problem." *Physica D* 42, 335-346.
- Harnad, S. (1991). "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem." *Minds and Machines* 1: 43-54.

- Harnad, S. (1992) “The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion.” *SIGART Bulletin* 3(4) (October 1992): 9 - 10
- Harnad, S. (1994) Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. *Artificial Life* 1(3): 1-10.
- Harnad, S. (2000). Minds, Machines and Turing: The Indistinguishability of Indistinguishables. *Journal of Logic, Language, and Information* 9: 425–445: 2000.
- Hoshika, S., Leal, N.A., Kim, M/J., Kim, M.S., Karalkar, Nilesh B., Kim, H.J., Bates, A.M., Watkins Jr., N.E., SantaLucia, H.A., Meyer, A.J., DasGupta, S., Piccirilli, J.A., Ellington, A.D., SantaLucia Jr., J., Georgiadis, M.M., Benner, S.A. (2019). “Hachimoji DNA and RNA: A Genetic System with Eight Building Blocks.” *Science* 363 (6429): 884–887.
- Jolliffe, D. and Farrington, D.P. (2006). “Development and Validation of the Basic Empathy Scale.” *Journal of Adolescence* 29: 89–611.
- Kahn, P., Kanda, T., Ishiguro, H., Freier, N.G., Severson, R.L., Gill, B.T., Ruckert, J.H. & Shen, S. (2012). ““Robovie, you’ll have to go into the closet now”: Children’s social and moral relationships with a humanoid robot.” *Developmental Psychology* 48(2): 303-314.
- Maibom, H.L. (2014). “Introduction: (Almost) Everything You Ever Wanted to Know About Empathy.” In *Empathy and Morality* (Heidi L Maibom, Ed.). Toronto: Oxford University Press. (pp. 1–40).
- Maibom, H.L. (2012). “The Many Faces of Empathy and Their Relation to Prosocial Action and Aggression Inhibition.” *WIREs Cogn. Sci.* 3: 253–263.
- Maibom, H.L. (2007). “The Presence of Others.” *Philosophical Studies*, 132(2): 161–190.
- Misselhorn, C. (2009). “Empathy with Inanimate Objects and the Uncanny Valley.” *Minds and Machines* 19: 345–359.
- Newell, A. and Simon, H. (1976). “Computer Science as Empirical Inquiry: Symbols and Search.” *Communications of the ACM* 19(3): pp. 113-126.
- Newman, M.H.A., Turing, A.M., Jefferson, G and Braithwaite, R.B. (1952). “Can Automatic Calculating Machines be Said to Think?” Radio Interview, original broadcast 14 and 23 January 1952. Turing Archives ref. B.6. Reprinted in Shieber, S. (Ed.) (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. (pp. 117-132). Cambridge, Mass.: The MIT Press.

- Nomura, T., Kanda, T. and Suzuki, T. (2006). “Experimental investigation into influence of negative attitudes toward robots on human–robot interaction.” *AI & Society* 20: 138–150.
- Nomura, T., Suzuki, T., Kanda, T. and Kato, K. (2006). “Measurement of negative attitudes toward robots.” *Interaction Studies* 7(3): 437–454.
- Nomura, T., Kanda, T., Suzuki, T. and Kato, K. (2008). “Prediction of Human Behavior in Human–Robot Interaction Using Psychological Scales for Anxiety and Negative Attitudes Toward Robots.” *IEEE Transactions on Robotics* 24(2): 442–451.
- Redstone, J. (2016). “Making Sense of Empathy with Sociable Robots: A New Look at the ‘Imaginative Perception of Emotion.’” In M. Nørskov (Ed.), *Social Robots: Boundaries, Potential, Challenges*. Farnham, UK: Ashgate.
- Rosenthal-von der Pütten, A.M., Krämer, N.C., Hoffmann, L., Sobieraj, S. and Eimler, S.C. (2012). “An Experimental Study on Emotional Reactions Towards a Robot.” *International Journal of Social Robotics* 5: 17–34.
- Rosenthal-von der Pütten, A.M., Schulte, F.P., Eimler, S.C., Hoffmann, L., Sobieraj, S., Maderwald, S., Krämer, N.C. & Brand, M. (2013). “Neural Correlates of Empathy Towards Robots.” *IEEE HRI 2013 Proceedings*: 215–216.
- Searle, J.R. (1980). “Minds, Brains and Programs.” *The Behavioral and Brain Sciences* 3: 417-457.
- Sharples, S., Cobb, S., Moody, A. & Wilson, J.R. (2008). “Virtual Reality Induced Symptoms and Effects (VRISE): Comparison of Head Mounted Display (HMD), Desktop and Projection Display Systems.” *Displays* 29: pp 58–69.
- Shieber, S. (Ed.). (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. Cambridge, Mass.: The MIT Press.
- Slater, M., Antley, A., Davison, A., Swapp, D., Guger, C., Barker, C., Pistrang, N. & Sanchez-Vives, M.V. (2006). “A Virtual Reprise of the Stanley Milgram Obedience Experiments.” *PLoS ONE* 1.1, e39.
- Sundman, J. (2003, February 26). Artificial Stupidity. Retrieved from https://www.salon.com/2003/02/26/loebner_part_one/
- Thagard, P. (2005). *Mind: An Introduction to Cognitive Science, 2nd Edition*. Cambridge: MIT Press.

- Turing, A.M. (1936). "On Computable Numbers, with an Application to the *Entscheidungsproblem*." *Proceedings of the London Mathematical Society* 2(42): 230–265.
- Turing, A.M. (1950). "Computing Machinery and Intelligence." *Mind* LIX(236): 433-60.
- Turing, A.M. (1951a). "Intelligent Machinery, a Heretical Theory." Lecture delivered to "51 Society" at Manchester, England c. 1951. Turing Archives ref. B.4. Reprinted in Shieber, Stuart (Ed.) (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. (pp. 105-109). Cambridge, Mass.: The MIT Press.
- Turing, A.M. (1951b). "Can Digital Computers Think?" Transcript of talk broadcast in BBC Third Programme, 15 May 1951. Turing Archives ref. B.5. Reprinted in Shieber, Stuart (Ed.) (2004). *The Turing Test: Verbal Behavior as the Hallmark of Intelligence*. (pp. 111-116). Cambridge, Mass.: The MIT Press.
- Weizenbaum, J. (1966). "ELIZA – A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9(1), 36-45.