The Questioning Turing Test

Nicola Damassino

PhD Philosophy

The University of Edinburgh

2019

Declaration

I declare that this thesis has been composed solely by myself and that it is the result of my own work, except where I indicate otherwise by reference or acknowledgement. This work has not been submitted for any other degree or professional qualification except as specified.

Lay Summary

Suppose you have an online, text-based chat with a stranger.

You can talk about whatever you want; and your interlocutor always replies something in an appropriate and convincing style, although it is not always cooperative, or even correct.

Suddenly, you ask yourself: is your interlocutor intelligent? Or, better, is it human?

This is a – simplified – Turing Test. Here, a human judge decides whether an unknown entity is human or machine, based on its interactions during a text-based conversation. However, evaluating the entity's ability to provide the correct responses, or probing its methodology in doing so, is not contemplated. This means that the entity can pass by producing uncooperative, parrot-fashion or even incorrect interactions, as long as they are human-like enough.

To prevent this, I propose a new version of the Turing Test: the Questioning Turing Test, where the unknown entity has to accomplish an enquiry by asking as few human-like questions as possible to a human judge. Here, three factors are evaluated: (i) the style of the entity's questions, (ii) the entity's ability to accomplish the enquiry, and (iii) the entity's strategy in doing so, in terms of the number of questions asked – the fewer, the better.

“*Il est encore plus facile de juger de l’esprit d’un homme par ses questions que par ses réponses.*”

(Pierre-Marc-Gaston, duc de Lévis, 1764–1830)

Contents

Abstract

The Turing Test (TT) is an experimental paradigm to test for intelligence, where an entity's intelligence is inferred from its ability, during a text-based conversation, to be recognized as a human by the human judge. The advantage of this paradigm is that it encourages alternative versions of the test to be designed; and it can include any field of human endeavour. However, it has two major problems: (i) it can be passed by an entity that produces uncooperative but human-like responses (Artificial Stupidity); and (ii) it is not sensitive to *how* the entity produces the conversation (Blockhead).

In light of these two problems, I propose a new version of the TT, the Questioning Turing Test (QTT). In the QTT, the task of the entity is not to hold a conversation, but to accomplish an enquiry with as few human-like questions as possible. The job of the human judge is to provide the answers and, like in the TT, to decide whether the entity is human or machine.

The QTT has the advantage of parametrising the entity along two further dimensions in addition to 'human-likeness': 'correctness', evaluating *if* the entity accomplishes the enquiry; and 'strategicness', evaluating *how well* the entity carries out the enquiry, in terms of the number of questions asked – the fewer, the better. Moreover, in the experimental design of the QTT, the test is not the enquiry *per se*, but rather the comparison between the performances of humans and machines. The results gained from the QTT show that its experimental design minimises false positives and negatives; and avoids both Artificial Stupidity and Blockhead.

Introduction

In "*Computing Machinery and Intelligence*" (1950) Turing proposes his approach to test for intelligence, the Turing Test (TT). According to the main interpretation of the TT, intelligence is inferred from the ability of a hidden entity to produce, during a text-based conversation, sufficiently human-like interactions for a human judge to decide that the entity is human. The article, which introduces the idea of thinking machines, that is, of machines able to do anything the human brain can, is regarded as a milestone of Artificial Intelligence research.

Turing, is worth noting, was not a cognitive scientist or an engineer, he was a mathematician and logician. So, what is the link between Mathematics and Artificial Intelligence? One obvious answer is that Artificial Intelligence involves computers, and using computers in those early days required a deep understanding of computer science, as well as the math involved. It is certainly true, but the answer fails to explain why Turing was not simply interested in building a mechanical calculator or even a computer in the sense we know today, but rather a "thinking machine". It's useful to make a brief historical excursus.

Turing, during his years at Cambridge, focused his study to Hilbert's problem of decidability. The results of Turing's research were published in the paper "*On Computable Numbers, with an Application to the Entscheidungsproblem*" (1936), where he developed one of the first definitions of computation and, more importantly, he came to the conclusion, along with Gödel (1931) and Church (1936), that computation cannot achieve everything. In the paper, Turing proposes the concept of *Turing machine*, which Church (1937) describes as follows:

> “a human calculator, provided with pencil and paper and explicit instructions, can be regarded as a kind of Turing machine.” (42)

It’s worth noting that a *Turing machine* is a purely theoretical device, and its components and actions (the paper tape, moving left and right, testing for the presence of a symbol, deleting a symbol and printing a symbol) are very basic ones. Turing also designed the hypothetical device that became known as *Universal Turing machine*. Church and Turing’s results inspired the so-called Church-Turing thesis, which formalises the definition of computable functions, stating that a function is effectively computable, that is, it can be mechanically computed by an algorithm, if and only if it is computable by a *Turing machine*. Because Turing’s results can be said to be more intuitive than Church’s, and because Turing explicitly uses machines is his arguments, the importance of “*On Computable Numbers…*” (1936) is highlighted in Copeland (2004), who argues that it can be regarded as the founding work of modern computer science.

There is a number of possible *Turing machines*, each able to carry out a different procedure, by virtue of having a certain program (or *table of behaviour*). So, whereas a *Turing machine* corresponds to a program, a *Universal Turing machine* corresponds to a computer, which can be fed different programs as data. Turing (1936), describing a *Universal Turing machine*, states:

> “It is possible to invent a single machine which can be used to compute any computable sequence.” (241)

In other words, any computer can be simulated by another computer, given that the second computer has the appropriate program. Since a *Universal Turing machine* can in principle do everything a human brain can do, it may be possible to hold that Turing's work not only shows what machines can or cannot do, but also what procedures the human brain can or cannot carry out (no matter the time and resources needed).

Turing identifies, among others, two types of machines with interesting implications on machine intelligence. One is proposed in "*Systems of Logic Based on Ordinals*" (1939), where Turing introduces the idea of a *Turing machine* connected to an oracle, which is able to compute incomputable functions. For this reason, the *oracle* would be way more powerful than any machine, since its working could not be purely mechanical. As Turing (1939) emphasises:

> "We shall not go any further into the nature of this oracle apart from saying that it cannot be a machine." (173)

The *oracle* is intended to introduce the conflict between 'intelligence' and 'infallibility', and to undermine the notion that a machine should never fail. As Turing (1947) holds:

> "[…] if a machine is expected to be infallible, it cannot also be intelligent. There are several mathematical theorems which say almost exactly that. But these theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility." (118)

The second type of machine is called *unorganised machine*. It is a machine with no hand-coded program except for one: a learning program, which enables it to acquire new knowledge and to develop new skills, just as a child's brain is able to do. The *unorganised machine*, or child machine, shows that Turing was not only interested in physical realisations and engineering problems (involving, for instance, finding a way to store more memory or to increase the speed of the machine), but also on the construction of "instruction tables" (that is, the programming of a machine). More than once he argues that it is possible to build a thinking machine by implementing it with "adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme." (Turing 1950, p. 442). But, as Turing (1946) clarifies, speed and storage are important factors insomuch as they are functional for the running of the system's programs. Moreover, he foresees that such programs would require a new domain expertise. As Turing (1946) holds, such programs will not be made by engineers, but rather they:

> "will have to be made up by mathematicians with computing experiences and perhaps a certain puzzle-solving ability. There will probably be a great deal of work to be done, for every known process has got to be translated into instruction table form at some stage." (391)

These remarks about the importance of software reflect the trajectory of computer science. Moreover, the insight that machines and the human brain can compute the same functions or, in other words, that they can *learn* to do the same things, led Turing to the idea of learning machines. As proof of this, Turing (1948) states:

> "It is pointed out that the potentialities of the human intelligence can only be realized if suitable education is provided. The investigation mainly centres round an analogous teaching process applied to machines. […] If one also decided on quite definite 'teaching policies' these could also be programmed into the machine. One would then allow the whole system to run for an appreciable period, and then break in as a kind of 'inspector of schools' and see what progress had been made. One might also be

> able to make some progress with unorganised machines." (428-432)

So, Turing did not limit his interest in processes that a machine can carry out mechanically. His ultimate goal was to build a device able to behave as the human brain does. He envisioned *unorganised machines* able to modify their own programs and simulate the cognitive development and the cognitive capabilities typical of humans.

In 1948, Turing and his associates develop the first chess engine ever, Turochamp. Two years later, Turing (1950) proposes the TT. The soundness of the TT as a test of intelligence has been constantly debated, and the number of arguments in support of it equals the number of rejections. In the thesis, I focus on the experimental weaknesses of the TT, as pointed out, among others, by Hernández-Orallo (2017), who argues that:

> "The standard Turing test is not a valid and reliable test for HLMI [Human Level Machine Intelligence]. […] the Turing test aims at a quality and not a quantity. Even if judges can give scores, in the end any score of humanness is meaningless." (p. 129)

The purpose of the thesis is to propose a new test, which I call the Questioning Turing Test (QTT), in order to improve the original experimental design of the TT.

I argue that the problem with the TT is that it doesn't care *if* or *how well* an entity produces a conversation, as long as its interactions are human-like enough. As a consequence, the TT attracts projects that concentrate on how best to fool the judges, or to produce frivolous exchanges that are only questionable exhibitions of intelligence. In the QTT, the hidden entity has to produce an enquiry rather than a conversation, asking questions to the human judge in order to accomplish the aim of the enquiry. The experimental design of the QTT has the advantage to parametrise the entity along two further dimensions in addition to 'human-likeness': what I call 'correctness', evaluating *if* the entity accomplishes the aim of the enquiry; and what I call 'strategicness', evaluating *how well* the entity produces the enquiry. The thesis is divided into three parts.

Part I is dedicated to the review of the literature related to the TT. In Ch. 1, I describe the experimental concept and discuss the experimental design proposed by Turing (1950), and I distinguish between the "Literal Interpretation", advocated by Sterrett (2000), and the "Standard Interpretation", advocated by Moor (2001). In Ch. 2, I introduce one of the two main difficulties with the test: Artificial Stupidity, that is, the potential exploitation of 'human-likeness' by the entity, as shown, for instance, by Eliza's conversational strategies (Weizenbaum, 1966). In Ch. 3, I discuss the other major difficulty with the test, the logical possibility of Blockhead, a humongous look-up table which is able to produce an appropriate response to whatever stimulus (Block, 1981).

In Part II, I present my proposal for a new version of the TT, called QTT. In Ch. 4, I consider two alternative versions of the TT that inspired the experimental design of the QTT: the MIST (Minimum Intelligent Signal Test) proposed by McKinstry (1997, 2009); and the FT (Feigenbaum Test) proposed by Feigenbaum (2003). I also consider the extended versions of the TT: the 3T (Total Turing Test), proposed by Harnad (1989, 1991, 2000) and the 4T (Truly Total Turing Test), proposed by Schweizer (1998, 2012a). In Ch. 5, I describe the QTT, the new test that I propose with the purpose to avoid both Artificial Stupidity and Blockhead. In Ch. 6, I consider the objections to the QTT, in particular, that the QTT is redundant and chauvinistic. In Ch. 7, I justify the importance that the questioning process plays in – at least – three disciplines related with intelligence: Developmental Psychology, Pedagogy and Epistemology, as argued, among others, by Hintikka (1985, 1999, 2007).

Finally, Part III is focused on my experiment. In Ch. 8, I describe my study involving the TT and the QTT; and in Ch. 9, I discuss the results.

Part I

1 The Turing Test

In Ch. 1, I present the TT, followed by the discussion of Turing's "*Computing Machinery and Intelligence*" (1950). Here I discuss the two main interpretations of the experimental design of the TT: the "Literal Interpretation" (Sterrett 2000, Traiger 2000), which I advocate; and the "Standard Interpretation" (Copeland 2000, Moor 2001). I also discuss the interpretation of Turing's conception of intelligence proposed by Proudfoot (2013). Furthermore, I analyse the dialogues Turing (1950) provides as examples of potential conversations during the test. I show the strategies Turing has in mind for the entity to pass the TT; and, following Fokker (2012) in the interpretation of the chess puzzle, I suggest that the TT is not intended to evaluate 'human-likeness' alone. I rather hold that, in the TT, 'human-likeness' is conflated with another dimension, that I proposed to evaluate independently: I call this dimension 'correctness', that is, the ability to provide the right response to a given question (or, more generally, the right output to a given input).

2 Artificial Stupidity

In Ch. 2, I discuss Artificial Stupidity. With Artificial Stupidity I refer to the set of strategies that can be used by the hidden entity (either machine or human) to exploit the judge's beliefs by holding an uncooperative and evasive, but human-like, conversation; in other words, by violating Grice's (1975) Cooperative Principle. The first and most well-known example of Artificial Stupidity is Eliza, a chatbot programmed by Weizanbaum (1966). My argument, according to which the TT can generate false positives or negatives because of Artificial Stupidity, does not imply that the hidden entity should be infallible. Therefore, I distinguish between Artificial Stupidity and Artificial Fallibility, and I claim that the experimental design of the TT should prevent entities that implement artificially stupid strategies from passing. In order to do so, as anticipated in Ch. 1, I propose to parametrise the hidden entity along a further dimension in addition to 'human-likeness': 'correctness'. It is worth noting that whereas Artificial Fallibility is compatible with both 'human-likeness' and 'correctness', Artificial Stupidity is compatible only with 'human-likeness', not with 'correctness'. To clarify this distinction, I show the Computer Game Bot TT, proposed by Hingston (2009), where Artificial Stupidity is used to exploit videogame players' beliefs and to enhance their recreational experience (fooling them into believing they are actually good at the game).

3 Blockhead

In Ch. 3, I review Block's "*Psychologism and Behaviourism*" (1981), and I describe the thought experiment which is known as Blockhead: an unimaginably huge pre-coded look-up table, or search tree, containing the appropriate verbal response to any possible verbal stimulus whatsoever. Blockhead undermines the validity of the TT showing that it can be passed by means of brute-force processes. In order to prevent Blockhead from passing the test, I propose a further dimension, in addition to 'human-likeness' and 'correctness', along which to parametrise the entity: 'strategicness', intended to evaluate *how well* the entity accomplishes the task set by the test. In order to evaluate 'strategicness', I propose to switch from the SISO setup (symbols in, symbols out), where a brute-force approach can always be successful, to the SOSI one (symbols out, symbols in), where a brute-force approach alone can never be successful. Whereas in a SISO test (TT) the task of the entity is to reply to the judge's questions, in a SOSI test (QTT) the task of the entity is to ask questions to the judge.

Part II

4 The Alternative Versions of the TT

In Ch. 4, I show two alternative versions of the TT which inspired the QTT: the Minimum Intelligent Signal Test (MIST), proposed by McKinstry (1997, 2009); and the Feigenbaum Test (FT), proposed by Feigenbaum (2003). The MIST is an automatable and quantitative test, where the hidden entity has to answer 20 yes/no subcognitive questions, the kind of questions that French (1990) argues to be the critical ones to show 'human-likeness'. The FT is a subject matter expert test, where the hidden entity has to answer expert questions about a certain domain, asked by a human judge, who is required to be an expert in that domain. I conclude that both the MIST and the FT can avoid Artificial Stupidity, by evaluating the replies given by the entity under scrutiny in terms of 'correctness'. However, they cannot avoid Blockhead, since they are SISO tests. I also discuss the extended versions of the TT: the Total Turing Test (3T), proposed by Harnad (1989, 1991, 2000); and the Truly Total Turing Test (4T), proposed by Schweizer (1998, 2012a).

5 The Questioning Turing Test

In Ch. 5, I describe the new version of the TT that I propose. The QTT is a SOSI test, where the hidden entity asks questions to the human judge, in order to accomplish the aim of an enquiry. The judge, in turn, has to decide (i) whether the entity is human or machine and (ii) whether the entity is able to accomplish the aim of the enquiry. The dimensions along which the entity is parametrised are three: 'human-likeness', ‘correctness’ and ‘strategicness’. ‘Human-likeness’ evaluates the entity’s ability to ask questions in a convincingly enough human-like fashion to be recognised as human by the judge. ‘Correctness’ evaluates the entity’s ability to accomplish the aim of the enquiry. And ‘strategicness’ evaluates the entity’s ability to show a *good* interrogative method, where *good* is defined in terms of the number of questions needed to accomplish the aim of the enquiry (the fewer the questions, the better the strategy). My claim is that the QTT, thanks to its experimental design, has the advantage to prevent both Artificial Stupidity and Blockhead from passing.

6 The Objections

In Ch. 6, I discuss the objections that can be raised against the QTT. The three main arguments are the following: (i) the QTT is redundant, for the TT can contain any verbal test whatsoever, QTT included; (ii) the QTT is chauvinistic, for only a good questioner can pass it; and (iii) the QTT cannot avoid Blockhead, for a questioning Blockhead would be able to ask any possible question.

7 The Questioning Process

In Ch. 7, I justify the importance that the questioning process plays in intelligence; and why the interrogative model should be preferred in testing for intelligence. I point out that – at least – three disciplines focus on the ability to ask good questions: (i) Developmental Psychology, specifically the relationship between questions and the cognitive development of individuals; (ii) Pedagogy, specifically the relationship between questions and learning; and (iii) Epistemology, specifically the relationship between questions and knowledge. I also discuss the work of Hintikka (1985, 1999, 2007) and in particular his Interrogative Game, intended to highlight the relationship between questions and deduction.

Part III

8 The Experiment

In Ch. 8, I describe the practical QTT that I designed in order to conduct my study and to show the advantages of the QTT over the TT. I recall the distinction between SISO (symbols in, symbols out) and SOSI (symbols out, symbols in); and I recall the two dimensions, in addition to 'human-likeness', along which the hidden entity is parametrised: 'correctness', intended to prevent Artificial Stupidity from passing by exploiting the judge's beliefs; and 'strategicness', intended to prevent Blockhead from passing by means of a brute-force approach. The experiment is divided into four phases. In the first phase, I conduct the original TT, where the hidden entity – either human or machine – is evaluated in terms of 'human-likeness'. In the second phase, I conduct the TT2, an alternative version of the TT where the hidden entity – either human or machine – is evaluated in terms of 'human-likeness' and 'correctness'. In the third phase, I conduct the QTT, where the hidden entity – either human or machine – is evaluated in terms of 'human-likeness', 'correctness' and 'strategicness'. And, finally, in the fourth phase, I conduct the Hybrid QTT, where the hidden entity is played by both a human and a machine, and the performance is evaluated in terms of 'human-likeness', 'correctness' and 'strategicness'.

9 The Results

In Ch. 9, I show and analyse the data gained from my study. First, the results of the original TT, which evaluates the hidden entity in terms of 'human-likeness' alone, confirm the potential unreliability of the TT's experimental design due to the false positives and negatives. Second, the results of the TT2 show that, by testing the entity in terms of both 'human-likeness' and 'correctness', the judge's biases are largely prevented: false positives are factored out, and false negatives are minimised. Third, the results of the QTT, where the hidden entity is parametrised along ‘human-likeness’, ‘correctness’ and ‘strategicness’, show that false positives and negatives are further reduced in comparison with the TT. Finally, the results of the Hybrid QTT, where the hidden entity is played by both a human and a machine, show the best performances and the best ratings in terms of ‘human-likeness’, ‘correctness’ and ‘strategicness’, outscoring the performances of both humans and machines alone in the previous tests.

Part I

Chapter 1
The Turing Test

Abstract. The first part of the thesis is dedicated to the discussion of the Turing Test (TT), the different interpretations that have been proposed and the difficulties with its experimental design. The TT is a procedure to evaluate whether a target entity can be considered intelligent, thanks exclusively to its conversational competency during a text-based conversation with a human judge. In this chapter, I present Turing's paper "*Computing Machinery and Intelligence*" (1950) and I review the two main interpretations of the TT, the "Literal Interpretation" (Sterrett 2000, Traiger 2000) and the "Standard Interpretation" (Copeland 2000, Moor 2001). I agree with Sterrett (2000) that the "Original Imitation Game" (OIG) provides the proper experimental design for the test, where the results are given by the comparison between the human's performance and the entity's performance; and with Traiger (2000) that the experimental design of the TT "invites generalization".

This chapter is dedicated to the discussion of Turing's procedure to evaluate whether a target entity can be attributed with intelligence. In his seminal paper "*Computing Machinery and Intelligence*" (1950), Turing proposes the Imitation Game (IG), renamed after its author the Turing Test (TT). In the TT, the property of intelligence is inferred from 'human-likeness', that is, the ability to be recognised as human by a human judge during a text-based conversation[1]. The TT has been at the centre of a lively debate in the last 70 years, which involves a philosophical and a practical problem: the former can be summarised by Turing's question "can machines think?" (and, by extension, by the

[1] With "text-based conversation" I mean a conversation carried out via text exchange, where not only are all visual cues unavailable, but also all cues from modalities such as sound, touch, smell, and so on.

question “what is it to think?”); the latter is focused on the project of building a machine that is able to pass the TT.

In the next section, I describe the Imitation Game, as presented by Turing (1950). In section 1.2, I discuss the “Literal Interpretation” and the “Standard Interpretation” of the TT. In section 1.3, I examine the role of the human judge in the TT. In section 1.4, I show the two dialogues Turing (1950) provides as examples of conversations during the TT. In section 1.5, I consider Turing’s conception of intelligence. And In section 1.6, I discuss French’s (1990) subcognitive argument.

1.1 Turing’s Test: The Imitation Game

Turing, in his seminal paper “*Computational Machinery and Intelligence*” (1950), asks the following question: “Can machines think?”[2] This question, Turing argues, is not a well-posed one, since it implies definitions that we do not have, and it should be reformulated to be “expressed in relatively unambiguous words.”[3] According to Turing, to answer the question we would need a definition of ‘thinking’ or ‘intelligence’, which is a problem. Turing, “instead of attempting such a definition [of intelligence],”[4] intends to reformulate in different terms the original question by proposing the Imitation Game. The IG, referred by Turing as a test in more than one occasion, and therefore renamed the TT, has been criticised by several authors. Hernández-Orallo (2017), among others, claims that:

> “The Standard Turing Test is not a valid and reliable test for HLMI [Human-level Machine Intelligence].” (129)

[2] Turing (1950, p. 433).
[3] *Ibidem*.
[4] *Ibidem*.

In this section, I describe the IG as presented by Turing (1950). To do so, I would like to begin by summarizing the first paragraph of "*Computational Machinery and Intelligence*" in current terminology, to clarify the scale and scope of Turing's project:

> Turing's purpose is to answer the question "can machines think?" by proposing a different method, since the original question is expressed in ambiguous terms which need strict definitions that are not readily available (especially the definition of intelligence). The goal of the IG, therefore, is not to provide a definition of intelligence whatsoever, but rather to provide a different way to answer the question whether a target entity can be said to be intelligent.

There is no evidence in Turing's literature that the IG is intended to provide a definition of (that is, to provide the logically necessary and sufficient conditions for) intelligence. There is, on the other hand, evidence that shows the contrary. In a BBC broadcast, when replying to Jefferson demanding what intelligence is, Turing (1952) explicitly admits:

> "I don't want to give a definition of thinking, but if I had to I should probably be unable to say anything more about it than that it was a sort of buzzing that went on inside my head. But I don't really see that we need to agree on a definition at all." (494)

So, according to Turing, the problem of evaluating whether an entity thinks[5] should be replaced with a new, more tractable problem, that is, whether the entity passes the IG. The IG, as described by Turing (1950), involves three participants per round, and two rounds in total. In every

---

[5] I use "to think" and "to show intelligence" interchangeably.

round there are two contestants (player A/M and B) and a judge (player C).

The first round is played by A (a man), B (a woman) and C (a human judge, no matter the sex). The two contestants and the judge are in separate rooms, and their conversation is carried on indirectly (in order to rule out any physical features that may simplify C's task). Most usually, this is done via text-based interactions, no matter the technological support: typewritten, teletyped, telegraphed, and so on.[6] So, it is worth noting, upgrading the test to the current technological level (by using, for instance, modern computers and chatrooms) does not affect its original design. The goal of C is to determine which of the two contestants is the woman and which is the man. More in detail, A must pretend to be a woman, by giving appropriate responses to C in order to make C misidentifying A as the woman; and B should help C into making the right identifications, namely that B is the woman and A is the man.

The second round is played in the exact same way and with the exact same task as the first, except that A is replaced by M. So, the second round of the IG involves M (a machine), B (a woman) and C (a human judge, no matter the sex). Just like in the first round, the goal of C is to determine which of the two contestants is the woman and which is the man. More in detail, M must pretend to be a woman, by giving appropriate responses to C in order to make C misidentifying M as the woman; and B should help C into making the right identifications, namely that B is the woman and M is the man.

The judge C asks questions to A/M and B by means of two terminals, simply labelled X and Y. At the end of each round, after a given period

[6] It is not important how the interactions are delivered, as long as they are indirect ones. As Turing (1950) points out, they could even "be repeated by an intermediary." (434)

of time and a series of unrestricted questions, C claims either: “X is A and Y is B or X is B and Y is A.”[7]

1.1.1 The TT’s Two Procedures

The IG, inspired by the parlour games, popular during the Victorian age, is intended to deal with the original question “can machines think?” by reformulating it into a new one. By testing A and B and then replacing A with M, Turing (1950) states, the new question is the following:

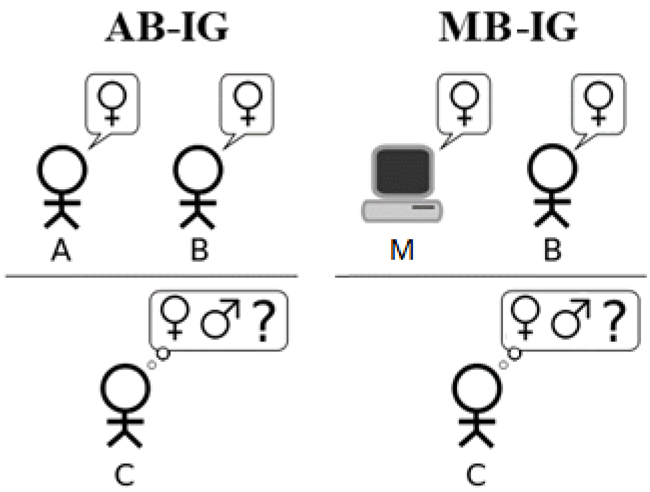
> “Will the interrogator decide wrongly *as often* when the game is played like this [between M and B] as he does when the game is played between a man and a woman [A and B]?” (434) [Italics added]

To clarify the IG, it is useful to distinguish between its two procedures: the AB-IG, that is, the first round of the IG; and the MB-IG, that is, the second round of the IG. The first refers to the IG played between two human contestants (a man and a woman); the second refers to the IG played between a machine and a human (a machine and a woman).

Along with Sterrett (2000), I hold that the results of the two procedures are not properly analysed if considered independently. Instead, the experimental design of the IG requires the outcomes of the two procedures to be compared. More specifically, the AB-IG provides the rate of C in misidentifying A or B, that is, the benchmark for the machine’s success in the MB-IG; and the MB-IG is based on that benchmark in order to determine whether C misidentifies M as B with the same frequency with which C misidentifies A as B. Both the pairs must carry out the same impersonation task (which, in the IG, involves the impersonation of a woman, but, as I will show, it can be modified

[7] Turing (1950, p. 433).

and adapted). If C misidentifies M in the MB-IG as frequently as C misidentifies A in the AB-IG, then the machine passes the IG, and it is attributed with intelligence as much as the man is. The following (fig. 1) is helpful to clarify the two procedures of the IG just distinguished:



(fig. 1)

There is a significant debate on what the IG exactly means: does Turing intend that a machine should successfully impersonate someone in order to be legitimately considered intelligent? Or does he rather intend that a machine should simply show a sufficient level of ‘human-likeness’ in order to convince C that it is human (and by extension intelligent), with no regard for impersonation? This is the long-standing debate between, respectively, the “Literal Interpretation” and the “Standard Interpretation” of the TT, which I discuss in section 1.2.

1.1.2 The Advantages of the TT

The advantage of the experimental conditions of the IG, Turing (1950) claims, is that they allow us to rule out a number of idiosyncrasies – of either humans or machines – from the test, making them irrelevant, as irrelevant for intelligence should be considered, for instance, the aesthetic features of machines or the cognitive limits of humans. As

Turing (1950) remarks, considering the problem of thinking machines from the IG perspective

> "has the advantage of drawing a fairly sharp line between the physical and intellectual capacities of a man." (434)

Moreover, the task set by the IG is, according to Turing (1950), suitable to include any field of human endeavour whatsoever, without the need of any physical demonstration. This highlights the versatility of the TT. However, it's worth noting, it is not *what* the participants talk about that helps C making the right identifications. Instead, it is *how* the participants talk about something. In other words, the participants can reply anything they want during the IG, but it's only thanks to the style of their replies that C can formulate a judgement about their nature. For instance, a machine being asked something like "are you a woman?" could reply in the following way: "yes I am a woman: as a matter of fact I have two eyes, a nose, a mouth, two arms, two legs..." and so on, listing all the human features that it is possible to find in an anatomy manual. However, it is not likely that C would misidentify that as a woman-like reply, not even a human-like one to begin with. On the contrary, if the reply were something like: "yeah, you might say so... even though sometimes my boyfriend insists that I cannot be human due to my snoring", C would have more troubles in deciding whether the reply comes from a woman or not (as well as whether it comes from a human or not). So, what really matters in the TT is the style of the interactions, not their contents. The dimension along which the entity is parametrised, and thanks to which C formulates her judgements, is what I call – conversational – 'human-likeness', that is the ability to be recognised as human by a human judge during a text-based conversation. As Turing (1950) puts it, the hidden entity's "best strategy is to try to provide answers that would naturally be given by a man."[8]

Turing's original question "can machines think?" could be rephrased in the following one:

> could a machine, during the text-based conversation in the MB-IG, be identified as human with at least the same frequency with which a human, during the text-based conversation in the AB-IG, is identified as human?

It is necessary to clarify, however, that Turing's goal is not to design an objective method to determine *with certainty* whether an entity is intelligent. In other words, he does not intend to provide the logically necessary and sufficient conditions for intelligence. Turing rather intends to propose the IG as an appropriate substitute for the question whether "machines can think". It is also worth noting that building a machine that could do well in the IG would not mean to create a machine replicating the human cognitive processes. As Turing (1950) argues:

> "May not machines carry out something which ought to be described as thinking but which is very different from what a man does? This objection is a very strong one, but at least we can say that if, nevertheless, a machine can be constructed to play the imitation game satisfactorily, we need not be troubled by this objection." (435)

Moreover, it is important to underline that Turing holds neither that machines could pass the IG and, consequently, that it's feasible to build a machine that can be considered intelligent; nor that machines could

[8] Turing (1950, p. 435).

not pass it, and, consequently, that it's not feasible to build a machine that can be considered intelligent. His purpose, as Turing (1952) explicitly writes, is simply to suggest that the new question posed by the IG is what should be considered:

> "I am not saying at present either that machines really could pass the test, or that they couldn't. My suggestion is just that this is the question we should discuss." (4)

1.1.3 The Interrogator in the TT

The proper conversation during the IG is described by Turing as *normal*; moreover, the proper judge, Turing states, should not be an expert, but rather an *average* person. With "normal conversation" and "average interrogator"[9] it is intended, respectively, a conversation that does not require a particular background or knowledge to be held and, consequently, a human judge who is not required to have such a background or knowledge. Most notably, Turing holds that the judge "should not be expert about machines"[10]; and Copeland holds that the judge should not be "expert about the human mind"[11] either. In other words, a computer scientist and a cognitive scientist would be banned from taking the test, in order to prevent, respectively, a machine from being unmasked due to its technical features, and a human from being recognized due to her cognitive characteristics. As Copeland (2000) remarks:

> "certain characteristic weaknesses in human reasoning – for example, a willingness in certain circumstances to assign a lower probability to a conjunct than to the conjunction, or the tendency to fail to take notice of certain disconfirming instances of conditional statements (weaknesses which are

[9] Turing (1950, p. 442).
[10] Turing (1952, p. 495).
[11] Copeland (2000, p. 525).

> easily detectable by Wason tests and the like) – could be used to unmask any computer not specifically programmed to reproduce these human foibles." (525)

1.1.4 The TT's Setups

Finally, the IG has been described in three distinct setups by Turing. First, the *parallel-paired* version, where a human judge engages text-based interactions with two hidden entities at the same time, either both human or one human and one machine[12], in order to decide which is which. Second, the "*viva voce*"[13] version, where a human judge engages text-based interactions one-to-one with a hidden entity in order to decide whether it is a human or a machine. And third, the *jury-based* version, where the hidden entities, either humans and machines, are "kept in a far away room and the jury are allowed to ask [them] questions"[14] in order to rate each of them as human or machine. In this work, I will focus on the *viva voce* version of the TT, discussing the *parallel-paired* version and discarding the *jury-based* version along with Copeland (2000), who argues that it

> "is open to a biassing effect, which disfavours the machine." (525)

The results from the *jury-based* tests reveal that there is a strong tendency among jurors to consider, for safety, humans as machines rather than the opposite. This bias in disfavour of the machine is called the Confederate Effect and, as Copeland (2000) puts it,

> "presumably this phenomenon is the result of a determination on the part of the jurors not to be fooled by a program." (*ibid.*)

---

[12] Turing (1950, p. 433).
[13] Turing (1950, p. 446).
[14] Turing (1952, p. 495).

Summing up, Turing (1950) introduces the IG in order to reformulate the question whether "machines can think" into the following one: could a machine do so well in holding a text-based conversation with an average human judge in order to be recognised as human with – at least – the same frequency with which a human can? Where 'average' is intended to factor out potential false positives or negatives (Proudfoot, 2013) by precluding computer scientists[15], or experts of the human mind[16], from playing the role of the judge. The IG, as described by Turing (1950) involves three participants: player A/M (a man or a machine); player B (a woman); and a human judge C (whose sex is irrelevant). And it is divided into two stages: the AB-IG, where the participants are both humans; and the MB-IG, where the machine challenges the human. Contestants A, B and M communicate with C in separate rooms and by indirect ways (e.g. text-based interactions on a computer chatroom). In the AB-IG, the goal of C is to decide which player is the man and which one is the woman; and the goal of both A and B is to be identified as the woman. In the MB-IG, player A is replaced by a machine M, and the game is played in the same way as the AB-IG. Turing does not intend to provide the logically necessary or sufficient conditions for intelligence, and the machine is not required to have the same processes which occur in the human brain. The test is intended to compare the rate of the judge's misidentifications when questioning a machine and a human with the rate of the judge's misidentifications when questioning two humans. In the next section, I discuss the two interpretations of the TT: the (i) Literal Interpretation, where both the AB-IG and the MB-IG are played; and the (ii) Standard Interpretation, where the AB-IG is skipped as merely introductory, and only the MB-IG is played.

[15] See Turing (1952).
[16] See Copeland (2000).

1.2 The “Literal” and “Standard” Interpretations

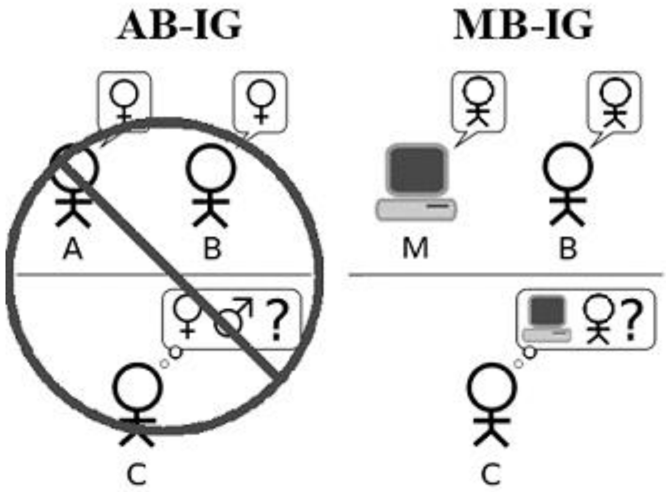
A critical passage in Turing (1950) is the second formulation of the IG, that is, when player A is replaced by a machine, which has

> “to play satisfactorily the part of A in the imitation game,
> the part of B being taken by a man.” (442)

This passage has generated two different interpretations, since player B, in the first formulation of the IG, is a woman, not a man.

The “Literal Interpretation” (LI) holds that what Turing proposes is a test made of two procedures: in the first, the contestants are two humans; in the second, the contestants are a machine and a human (see fig. 1). And the results are given by the comparison of their performances in both the procedures. Or better, by the comparison of the frequency with which the judge makes the wrong identifications in the AB-IG, with the frequency with which the judge makes the wrong identifications in the MB-IG.

The “Standard Interpretation” (SI) holds that what Turing intends is to test the hidden entity for its ability to be human-like enough to be indistinguishable from a real human: according to the SI, only the machine vs human test is important, whereas the human vs human one is just introductory. The following (fig. 2) depicts the Standard TT (STT):



(fig. 2)

1.2.1 Literal Interpretation

Sterrett (2000) and Traiger (2000) both argue in favour of the LI, holding that the two procedures of the IG (AB-IG and MB-IG) should be kept distinct, and that it is necessary to run both of them in order to conduct the test properly. The main idea, supported by Sterrett (2000), is that:

> "the advantage of Turing's first formulation of the test [OIG] is that it provides a context in which the computer and the man are put on a more equal footing: both the computer and the man will need to critically evaluate their responses, and fabricate appropriate ones that are based on vicarious experiences of womanhood." (552)

This, however, does not imply that 'womanness' or femininity are the properties to evaluate when testing for intelligence. In other words, cross-gendering is not a necessary condition to test for intelligence. As Sterrett (2002a) clarifies:

> "I stated that gender impersonation and discrimination were inessential to the test, while explaining why it was a good task to employ in a test of intelligence: "The significance of the cross-gendering in Turing's Original Imitation Game Test lies in the self-conscious critique of one's ingrained responses it requires. And, that the critique has two aspects: recognizing and suppressing an inappropriate response, and fabricating an appropriate one" (Sterrett, 2000, p. 551)." (132)

Traiger (2000) argues in favour of the LI for two reasons. The first is that the LI-IG (that is, the OIG) can provide a less biased basis due to the kind of questions that the judge would ask. As he holds, given the experimental design of the LI-IG,

> "the kinds of questions posed by the interrogator will be those that would typically be posed to human participants. The interrogator in the Imitation Game will not begin with the hypothesis that one participant is a computer, and hence the computer participant, like its human counterpart, can devote its energies to answering the same kinds of questions that would typically be posed to a human." (569)

The second is that, in the LI-IG, it is not possible for the contestants to exploit the judge's beliefs. On the other hand, in the SI-IG, Traiger (2000) holds,

> "the computer and the human competitor may exploit beliefs about machines to try to trick or help the interrogator." (566)

1.2.2 Standard Interpretation

Rejecting the LI, Moor (2001),Copeland (2000) and Piccinini (2000) argue in favour of the SI. The reason is that Turing never points out femininity – or gender in general – to play a role in intelligence, as argued by Copeland (2000)[17] and Moor (2001)[18]. And Piccinini (2000) claims that, throughout Turing's works, it is possible to find "plenty of evidence that the standard reading of his rules is correct."[19] One piece of evidence in support of the SI, Piccinini (2000) argues, is found in Turing (1950), when he explicitly writes that there would be

> "little point in trying to make a thinking machine *more human* by dressing it up in […] artificial flesh." (434) [Italics added]

According to Piccinini (2000), the quoted passage above not only tells us that Turing does not consider aesthetic features to be necessary for intelligence, but also that he does not intend to make a thinking machine more "manly" or "womanly", just more "human". Turing does not mention, as Moor (2000) underlines, femininity or masculinity, but rather humanity; and so, as Moor insists, the IG should be intended to test for 'human-likeness', irrespective of gender. Moor (2001), however, goes as far as claiming that the AB-IG,

> "is at most an intermediary step toward the more generalized game involving human imitation." (79)

---

[17] Copeland (2000) holds that the TT is intended "[…] to determine whether or not a computer can 'imitate the brain'. It seems unlikely, therefore, that Turing's intention in 1950 was to endorse only the female-impersonator form of the test, or that he saw himself as describing different tests […]." (526)

[18] See Moor (2001): "Turing continues throughout the rest of his paper to emphasize humanity not femininity." (78)

[19] See Piccinini (2000, p. 580).

1.2.3 Advantages of the LI over the SI

I support the view according to which both the AB-IG and the MB-IG are necessary procedures in order to test the hidden entity properly, that is, in order to run the TT properly. I agree with Sterrett (2000) and Proudfoot (2013) in holding that the AB-IG scores the MB-IG and provides a benchmark for the machine's success. I also agree with Traiger (2000), who holds that the description of the IG in Turing (1950) implicitly "invites generalization"[20]. That is, when Turing specifies that the goal of the judge is to claim either "X is A and Y is B or X is B and Y is A"[21], those abbreviations are not intended for brevity only. As Traiger (2000) writes:

> "As many commentators have noted, there's nothing special about the game determining whether someone is a male or a female. [...] If this is right, "A" and "B" could be placeholders for whatever characteristics may be used in different versions of the game." (565)

The IG, in other words, can be played in different variations, involving other characteristics rather than those related to gender. For instance, instead of a man and a woman, the IG could be played between an eastern and a western woman, a young boy and an old man, a native speaker and a non-native one, and so on. This is compatible with Sterrett (2000), who argues:

> "[…] cross-gendering is not essential to the test; some other aspect of human life might well serve in constructing a test that requires such self-conscious critique of one's ingrained responses." (pp. 550-551)

[20] Traiger (2000, p. 565).
[21] Turing (1950, p. 433).

According to Sterrett (2002a), the essential characteristic of the TT, which should be preserved in any experimental variations that can be designed, is that it is a nested test:

> "One distinctive feature is the "nested algorithm" structure of the test. The Original Imitation Game is played repeatedly in the course of a single test, and is nested inside the test. The test is a practical test, which, if passed, would be evidence for regarding a machine's behavior as intelligent. The game is not the test; the game is nested *inside* the test, as a DO or DO WHILE loop is nested in an algorithm." (132)

So, whereas in the STT the machine is penalised, and the role of the judge is vested with heavy responsibilities, the OIG, as Sterrett (2002a) points out:

> "(1) […] permits the result that the machine does better than the man, (2) […] tends to screen off lack of interrogator skill, and (3) both man and machine are required to impersonate. The machine's performance is not directly compared to the man's, but their rates of successfully impersonating against a real woman candidate are compared." (131)

Against the LI, Moor (2001) holds the OIG to be redundant, since "the aspects of intelligence that Sterrett identifies as important to test can be tested in the standard game."[22] However, as Sterrett (2002a) observes, even if the STT can test some of the aspect that are important in the OIG, it does not follow that the STT is a better test than the OIG. On the contrary, there are two crucial differences between the OIG and the

[22] See Moor (2001, p. 80).

STT: one is that the interrogator's gullibility plays a bigger role in the STT than in the OIG; and the other is that in the STT, only the machine has a task, whereas in the OIG both the machine and the human have one. Therefore, as Sterrett argue, the OIG does have a normative advantage[23] over the STT.

The role of the judge in the STT is summed up by Sterrett (2002a) as follows:

> "[…] The "Standard Turing Test" is just too sensitive to the skill of the interrogator to even be regarded as a test." (135)

In other words, in the experimental design of the STT, the outcome heavily depends on how good the interrogator is at discriminating the entities[24]. This is not the case in the OIG, and the reason is that the task involved in the STT and in the OIG is not the same. As Sterrett (2002b) specifies:

> "In 'the original imitation game test', both the man and the computer are called upon to impersonate. What the test results reject is their relative success in achieving this goal. In the standard Turing test, the man is not called upon to do

[23] As Sterrett (2002b) clarifies, the differences between STT and OIG are (i) qualitative, since "[…] the standard Turing test is far more sensitive to the interrogator's skill." (44); and (ii) quantitative, since in the OIG "[…] there is nothing in the structure of the game to prevent a machine scoring higher than a man: […] here there is an independent measure of success other than the interrogator's direct comparisons of the responses of the man and the machine, because there is a task at which each can either fail or succeed, independently of the other. […] The standard Turing test does not even admit of such a result; there, the responses of the man and the machine are directly compared and only the machine is judged as to having exhibited a requisite skill or having failed to do so" (43).

[24] See Sterrett (2002b): "Sensitivity of the results of a test to the skill of the interrogator running it is hardly a virtue of any test. […] In 'the original imitation game test' the machine's intelligence is being tested by comparing the frequency with which it succeeds in causing the interrogator to wrongly identify it as the woman with the frequency with which a man succeeds at doing the same. Thus, C's skill level affects both frequencies. More fundamentally, the differences between the two tests are due to this: In `the original imitation game test', unlike in the standard Turing test, scoring as a thinker does not amount to simply being taken for one by a human judge." (44)

> anything very novel, whereas the computer must converse under the pretence that it is human! In contrast, `the original imitation game test' compares the abilities of man and machine to do something that requires resourcefulness of each of them. […] for both the man and the machine, the task set has these aspects: recognizing an inappropriate response, being able to override the habitual response, and being able to fabricate and replace it with an appropriate response." (44)

Finally, as Sterrett (2000) concludes, the relevance of the OIG over the STT lies not just in the different experimental design, but in how intelligence is characterised:

> "the importance of the first formulation [OIG] lies in the characterization of intelligence it yields. If we reflect on how the Original Imitation Game Test manages to succeed as an empirical, behavior-based test that employs comparison with a human's linguistic performance in constructing a criterion for evaluation, yet does not make mere indistinguishability from a human's linguistic performance the criterion, we see it is because it takes a longer view of intelligence then linguistic competence. In short: that intelligence lies, not in the having of cognitive habits developed in learning to converse, but in the exercise of the intellectual powers required to recognize, evaluate, and, when called for, override them." (558)

Summing up, I agree with Sterrett (2000) that both the AB-IG and the MB-IG are necessary for running the TT properly; and with Traiger (2000) that the TT is intended to "invite generalization". I reject Moor's (2001), Piccinini's (2000) and Copeland's (2000) perspective that the

AB-IG is just an intermediary step (with the only purpose to introduce the test to the reader). I hold that the human vs human game is rather intended to provide the benchmark with which the machine's performance will be scored in the machine vs human game. As Proudfoot (2013) puts it:

> "Turing used the man-imitates-woman game to score the computer-imitates-human game: a machine does well in the latter game when the interrogator in that game is fooled no less frequently than the interrogator in the former game." (395)

1.3 The Role of the Human Judge

In this section, I discuss the role of the human judge in the TT, who makes a decision about the nature of a target entity based on indirect interactions during a text-based conversation.

Turing (1950) argues that a conversational test of intelligence allows the experimenter to include almost any field of human endeavour, without penalising the inabilities of the machine (such as the inability to be hungry) or those of the human (such as the inability to make huge calculations). The experimental design of the TT, Turing holds, "make these disabilities irrelevant."[25] In other words, the TT discourages chauvinism.

Moreover, a test of intelligence where the outcome is decided by the human judge, and not objectively obtained, has no presumption to provide the logically necessary conditions for intelligence[26]. And, since

[25] Turing (1950, p. 435).
[26] Shieber (2004, p. 270).

passing the TT is not a necessary condition for intelligence, failing the TT does not rule out intelligence[27]. According to this reading, however, the TT should still be considered to provide the logically sufficient conditions for intelligence. This is compatible with Turing (1950), who seems to reject the idea that imitation is the adequate metric for intelligence. He argues that an optimal cognitive strategy for a machine "may possibly be something other than imitation of the behaviour of a man"[28], where it is assumed that the best strategy is "to try to provide answers that would naturally be given by a man."[29] And it is compatible with Cowen & Dawson (2009) as well, who point out that a further proof that Turing's intention is not to claim that the appropriate metric for intelligence is imitation is that, as Turing holds, lacking the ability to imitate does not mean lacking intelligence:

> "Imitation and intelligence are simply not the same and lack of imitation, or failure to be indistinguishable, does not mean lack of intelligence—or lack of humanity." (6)

So, given the experimental design of the TT, it is often claimed that passing it would provide a proof of the presence of intelligence, whereas failing it would not provide a proof of its absence. The problem with this reading, is that it is not that obvious that the TT is intended to provide the logically sufficient conditions for intelligence. The reason is that the outcome of the test does not depend solely on the entity's performance, but also on the judge's impression. I will discuss this in more detail below (see section 1.5).

---

[27] See Dennett (2004). See also Schweizer (1998): "I view such tests as dealing with *evidence* for intelligence, but I do not view them as constitutive or definitional. In particular, this means that their value is mainly for drawing positive rather than negative conclusions, and hence *failure* on such tests is not tantamount to lacking intelligence (or a mind); rather, the system has simply failed to exhibit the 'canonical' signs. The main role of behavior is inductive or evidential, and so behavioral tests do not provide a necessary condition nor a reductive definition. It then follows that failure to exhibit the relevant sort of behavior warrants no conclusion whatever." (264)

[28] Turing (1950, p. 435).

[29] *Ibidem*.

1.3.1 Problems with the judge

In the TT, the human judge plays a crucial role, for she is the only one who makes a decision about the nature of the entity under scrutiny. However, this decision is based on the judge's personal beliefs when evaluating a target entity. This leads the TT to provide potentially biased results, both in the LI-IG (where the verdict is given by the rate of the judge's misidentifications) and the SI-IG (where the verdict is given directly by the judge's decision): it could be the case where the judge always misidentifies the target entity due to her personal beliefs. In other words, the experimental design of the original TT, either LI-IG or SI-IG, is not screened out from the personal beliefs that potentially lead the judge to misidentifies the machine as human (Eliza Effect) or the human as machine (Confederate Effect).

Another important aspect of the experimental design of the TT is how the judge is to be chosen. This is still a widely debated topic: Hayes & Ford (1995), for instance, point out that:

> "the imitation game conditions say nothing about the judge, but the success of the game depends crucially on how clever, knowledgeable, and insightful the judge is." (973)

Turing does not require the judge to be particularly "clever, knowledgeable or insightful": on the contrary, as discussed above, the judge should simply be an average human (while computer or cognitive scientists should be banned).

Traiger (2000) argues that, on the one hand, in the OIG the judge does not know that one of the human participants is replaced with a machine. In the STT, on the other hand, the judge is aware that the interlocutor can be either a human or a machine. By trying to determine which is which, “the judge brings to the game his or her beliefs about computers and machine intelligence.”[30] Consequently, in the STT, the conversation is no longer held as intended, that is, the judge does not behave in the same way as she would with other human fellows. Unlike the judge in the OIG, the judge in the STT will rather engage the target entity in a difficult and odd conversation, made of paradoxical questions and semantic traps in order to unmask any potential machine[31]. The STT is responsible for giving importance to those very factors that the IG was intended to rule out from the judge’s judgement. As Traiger (2000) writes:

> “Interrogators setting out to unmask a computer will mobilize and rely on their own conceptions of what computers are and what they can and can’t do, just as the computer program may successfully exploit those very beliefs in order to appear human-like. And this makes the test dependent on such conceptions of computation and mechanism, precisely the uninteresting question bypassed by Turing’s Imitation Game.” (570)

1.3.2 The unaware judge

Traiger (2000) agrees about the role of the judge in the TT with Mauldin (1994), who proposes the Unsuspecting Turing Test (where the judge is not aware that the target entity might be a machine) as the proper setup in which the test should be run. In the UTT, the judge is not aware that the target entity might by a machine, and it is inspired by TinyMUD (a sandbox videogame descendent of MUD, which stands for Multi-User Dungeon), developed by James Aspnes in 1989. Mauldin’s project was to program an automated player able to independently explore the text-based world of TinyMUD. The resulting program, called

[30] Traiger (2000, p. 565).
[31] Traiger (2000, p. 569).

ChatterBot, was observed to be often taken for a human by other humans players. As Mauldin (1994) writes:

> "The ChatterBot succeeds in the Tiny-Mud world because it is an *unsuspecting Turing test*, meaning that the players assume everyone else playing is a person, and will give the ChatterBot the benefit of the doubt until it makes a major gaffe." (17)

Among Mauldin's programs there are Gloria, focused on gradually accreting more and more linguistic ability; and Julia, able not only to become a more capable conversational agent, but also to assume duties in the TinyMUD text-based world, such as tour guide, information assistant, note-taker, message-relayer, and so on. Mauldin's TinyMUD bots were able to pass the UTT because of the unaware judges, who had no reason to suspect that one of the other players was a program rather than a human, and so they were more polite and less inclined to ask probing questions.

1.4 Turing's Dialogues

In this section, I consider the two dialogues Turing provides (1950) as examples of conversations occurring during the TT. The setup in which the dialogues are presented is the *viva voce*, that is, as Turing (1950) specifies, when the test is conducted "with the player B omitted, frequently used in practice."[32] Here I consider each question and reply of the dialogue, and I discuss a few remarks that can be inferred from them.

[32] Turing (1950, p. 446).

1.4.1 First Dialogue

The first dialogue that Turing (1950) provides is the following:

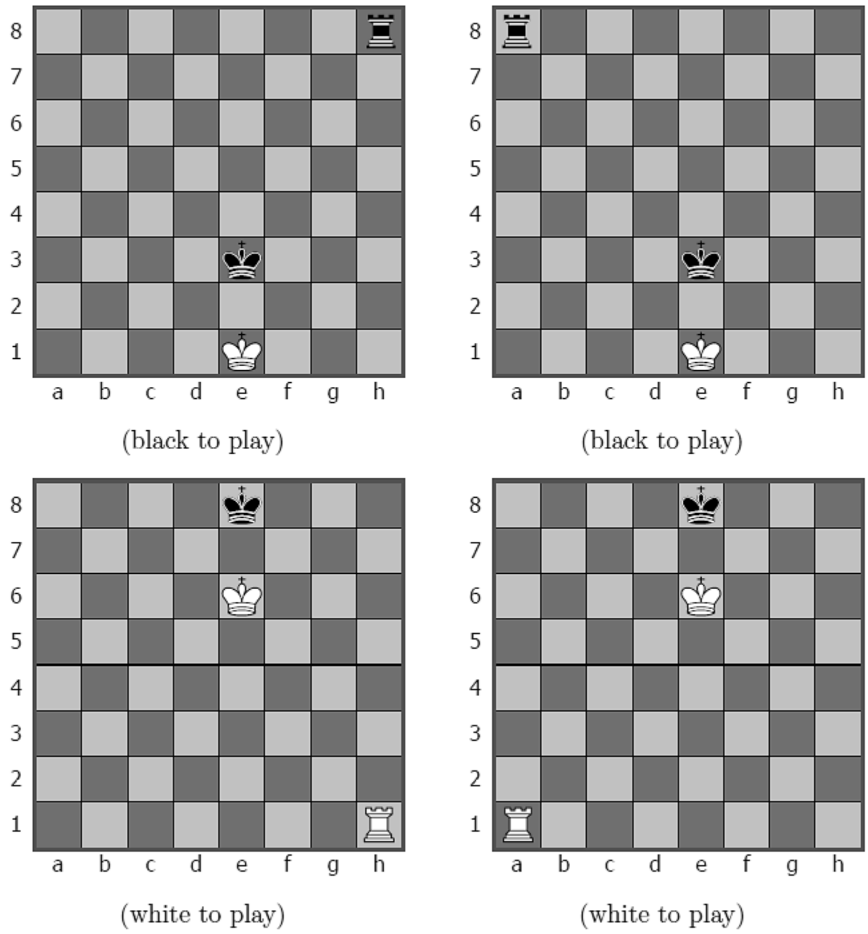
> "**Q**: Please write me a sonnet on the subject of the Forth Bridge.
>
> **A**: Count me out on this one. I never could write poetry.
>
> **Q**: Add 34957 to 70764
>
> **A**: (Pause about 30 seconds and then give as answer) 105621.
>
> **Q**: Do you play chess?
>
> **A**: Yes.
>
> **Q**: I have K at my Kl, and no other pieces. You have only K at K6 and R at Rl. It is your move. What do you play?
>
> **A**: (After a pause of 15 seconds) R-R8 mate." (435)

The first dialogue unveils a number of strategies that the hidden entity can use to pass the TT. In the first line, the judge asks the entity to write a sonnet, and the entity replies that it cannot write poetry. The answer has a convincingly human-like form, and it is also plausible that a human would reply in that way if asked to write offhand a poem. This first entry already reveals a potential exploit: the entity can evade or refuse any of the judge's interactions, as long as in a human-like fashion. I call this Artificial Stupidity (which I discuss in Ch. 2).

The second request of the judge is to perform an arithmetical addition. There are two different strategies used in the entity's reply. The first is that, as Turing explicitly indicates, the machine waits 30 seconds before giving the answer. The lesson here is that, no matter how fast the entity (human or machine) can make additions, it should delay its reply. More in general, an entity which instantly types difficult or long replies would be easily unmasked (a typist, for instance, may have problem if she types faster than an average person). This case also shows the importance of having the average human performance as a benchmark, in order to rule out exceptions like, for instance, people affected by autism (that might be extremely fast at calculating numbers and, therefore, misidentified as machines). The second strategy is even more subtle. The answer given by the entity is not the correct one: the exact result is 105.721, and not 105.621. However, it would be much more plausible for a human to give a slightly wrong result. So, the entity (either human or machine), given the experimental design of the TT, where intelligence is inferred from 'human-likeness', should adapt to the replies that an average human would give, in order not to be unmasked too easily or misidentified. In other words, the entity should have good strategies to avoid revealing its mechanic nature or its extraordinary skills (e.g. by delaying its replies and making mistakes). I refer to this as Artificial Fallibility (which I discuss in Ch. 2).

The final case that Turing provides in the first dialogue is a chess puzzle. Fokker (2012) analyses this case, and he draws some interesting conclusions. The judge provides the position of the three pieces on the board, two kings and one rook, and asks for the entity's move. The case, however, is more subtle and more interesting than it looks at first, as Turing sets aside something for the alert reader. The chess case, *per se*, is not a problematic one: it is not difficult to program a machine to play chess at a very expert level. So why Turing chooses such a simple problem, that even in the 50s would have been hardly challenging for a machine? It is not for brevity: on the contrary, the chess case represents the most difficult challenge in the first dialogue, since it requires access to common-sense understanding. First of all, the chess case is presented with the – now obsolete – Descriptive Notation, widely used until the late 20th century, but then replaced by the Algebraic Notation. And Turing does not even use the Descriptive Notation rigorously; he rather uses a mixed style: "K *at my* K1", "K *at* K6", "R *at* R1", "no other pieces”[33]. Thus, the entity needs to deal with some natural language processing, in order to gain the right data from the question (for instance, by distinguishing the formal language, that is the Descriptive Notation, from the other unusual phrasing). Moreover, the Descriptive Notation, unlike the Algebraic one, needs to be disambiguated: “R at R1” means that the rook could be the left one as well as the right one; in both cases, however, the entity’s reply “R-R8” is the correct one, as moving the rook, be it the left one or the right one, to R8 leads to checkmate. Even more important, in the Descriptive Notation the positions of the pieces are not fixed, but relative to the player: if the judge is white and the entity is black, “K at K6” would not work if it meant “King in K6”: it rather means “King in K3” (Ke3); on the other hand, if the judge is black and the entity is white, than “K at my K1” means “King in K8” (Ke8). This information is not provided, and it should be inferred by the entity (as shown in fig. 2).

[33] Turing (1950, p. 435).



(fig. 2)[34]

Moreover, the judge simply asks the entity "what do you play?"[35], and the entity is tacitly assumed to make a winning move, above all because it is a mate-in-one chess puzzle. But it is worth noting that it's a common-sense assumption that the move in the chess puzzle should be a winning one. As Halpern (2006) observes:

> "Turing's sample dialogue awards the computer just that property that programmers have never been able to give their computers: common sense." (46)

[34] Fokker (2012, p. 95).
[35] Turing (1950, p. 435).

In other words, it would not be that impressive if the entity could solve the problem; and the entity indeed solves it, by replying “R-R8”[36], after a pause of 15 seconds – for the same reason as the delayed reply in the arithmetical case. Turing’s example shows an entity that is able to understand natural language and deal with common-sense implications. In the chess case, the real task of the entity is not to solve the problem, but rather to show understanding. As Fokker (2012) puts:

> “the mere parsing of the questions is far more difficult to program than the actual chess problem solver.” (96)

I argue that the chess puzzle, as analysed by Fokker (2012), shows a major problem in the experimental design of the TT: the conflation between ‘human-likeness’, that is, the ability to be indistinguishable from a human being, and what I call ‘correctness’, that is, the ability to provide the right verbal response to a verbal stimulus. As I discuss in Ch. 2, the conflation of these dimensions weakens the TT, making it vulnerable to Artificial Stupidity: since the hidden entity needs to give human-like replies to be considered human, it does not matter if the replies are incorrect or even not pertinent at all. As long as those replies have a human-like style, the entity may be recognised as human and pass the TT, even by giving only wrong and uncooperative replies. In order to prevent this potential exploit of the TT, I propose to keep ‘human-likeness’ and ‘correctness’ independent dimensions along which to parametrise the entity. I will argue this proposal more extensively in the next chapters.

1.4.2 Second Dialogue

The second dialogue provided by Turing (1950) is the following:

[36] *Ibidem*.

> “**Interrogator**: In the first line of your sonnet which reads ‘Shall I compare thee to a summer’s day’, would not ‘a spring day’ do as well or better?
> **Witness**: It wouldn’t scan.
> **Interrogator**: How about ‘a winter’s day’ That would scan all right.
> **Witness**: Yes, but nobody wants to be compared to a winter’s day.
> **Interrogator**: Would you say Mr. Pickwick reminded you of Christmas?
> **Witness**: In a way.
> **Interrogator**: Yet Christmas is a winter’s day, and I do not think Mr. Pickwick would mind the comparison.
> **Witness**: I don’t think you’re serious. By a winter’s day one means a typical winter’s day, rather than a special one like Christmas.” (446)

The second dialogue is more explicitly intended to deal with discourse representation and common-sense understanding. The dialogue is intended as a reply to Jefferson (1949), who writes:

> “Not until a machine can write a sonnet or compose a concerto [...] could we agree that machine equals brain.” (1110)

Jefferson’s argument supports the view that machines cannot be said to think until they are intentional agents, and not acting merely “by the chance fall of symbols”[37]. Turing (1950) calls it the Argument from Consciousness, and rejects it due to its solipsistic consequence: how can we know that our fellow creatures act intentionally and not merely by chance? In other words, how can we know that other humans, except

[37] Turing (1950, p. 445).

ourselves, think? The dialogue is intended to reply to Jefferson's remark by providing a discussion about a poem between an entity and the judge.

The questions asked by the judge involve the prosody and the reasons why the entity uses a particular metaphorical dictionary, in order to unveil some human poetic sensibility. Turing picks one of the most praised sonnets by Shakespeare, Sonnet 18.[38] But what would Jefferson say, as Turing asks, "if the sonnet writing machine was able to answer like this in the *viva voce*?"[39] Would it be possible to describe it just an assembly of words, "an easy contrivance?"[40] It's safe to assume that here Turing implies that originality and creativity are rather problematic parameters, anticipating Lady Lovelace's objection, according to which "a machine can never do something new"[41].

I argue that the judge in the second dialogue provided by Turing is a poetry expert. It seems reasonable to say that the judge is required to be a poetry expert in order to test a sonnet-writing entity properly, since any judge could ask the entity to produce a poem, but only one with certain poetry expertise could evaluate the poem properly. And, moreover, another clue that might confirm the judge to be a literature expert, is that she mentions another literature character, namely Mr. Pickwick, and asks the entity a question the answer of which requires knowing "The Pickwick Papers" by Dickens. Because of Turing's suggestion of a test involving a "sonnet-writing machine" [42] and the dialogue above designed as a poetical version of the TT, I argue that Turing, as claimed by Traiger (2000), not only invites generalization in the TT, in terms of the task set by the test, but he also allows to run the test with experimental design variations, for instance, to run the TT as

[38] The sonnet provides a clear example of fine poetry that anyone could recognise as such. Interestingly enough, this could be a reference of Turing's private life, since Sonnet 18 can be interpreted, as the modern critics do, as "the articulation of same-sex love." (Cohen, 2008)
[39] Turing (1950, p. 446).
[40] Turing (1950, p. 447).
[41] Turing (1950, p. 450).
[42] Turing (1950, p. 446).

an Expert System test (allowing the judges to be expert in a certain field). So, according to this interpretation, the TT can be run as an Expert System test, in which the judge should not be just an average person, but rather an expert in a certain domain; and the conversation would not be a normal one, but rather specifically focused on that domain. If this is true, Turing anticipated the Feigenbaum Test (which I discuss in Ch. 4), proposed by Feigenbaum (2003), in which Expert Systems have to show advanced knowledge in a certain domain, by answering questions from a human expert in that domain, and by justifying the answers. An example of an Expert System TT such as the poetical TT, would thus involve a sonnet-writing entity and a poet judge, and the goal of the entity would be to "impress" the poet judge, by producing a good poem and by explaining it. Similarly to the poetical TT, it is possible to design a number of different tests: an example would be, following Jefferson (1949) in the quotation above, the musical TT, involving a concerto-composing entity and a musician. This, I hold, confirms Traiger's (2000) view discussed in section 1.3, according to which the IG is intended to invite generalization, and the task can be modified to infer intelligence from a number of different characteristics and abilities (not necessarily impersonation).

Summing up, I argue that Turing (1950) makes a few implicit and underestimated remarks when he provides the two dialogues as examples of conversations occurring during the TT. The first remark, made in the first dialogue, is that (i) the target entity, in order to pass the TT, should be as fallible as a human (for instance, it should reply in a delayed and, from time to time, wrong way to certain kind of questions). The second remark, made in the first dialogue, is that, in order to pass the TT, (ii) the target entity should have access to a common-sense knowledge base, and it should be parametrised not only along the dimension of 'human-likeness', but also along the dimension of 'correctness'. The third and the fourth remarks, made in the second dialogue, are (iii) that the experimental design of TT is intended to invite generalization and modifications; and (iv) that it allows more specific versions of the test to be played with experts, where ‘correctness’ is evaluated along ‘human-likeness’ (anticipating the Feigenbaum Test, which I describe in Ch. 4). Alternative versions of the test could be the Poetry TT or the Music TT (involving, for instance, Botpoet[43], a bot able to produce haiku poems; and Bachbot[44], a bot able to produce music inspired by Bach).

1.5 Turing’s Conception of Intelligence

In this section, I present Turing’s conception of intelligence, as Turing discusses and as Proudfoot (2013) analyses. Here, I address the following question: why did Turing choose to keep the judge an average person, with the risk for the TT to get biased results?

In “*Intelligent Machinery*” (1948), Turing explicitly describes intelligence as an “emotional concept”[45], that is, a concept the applications of which are

> “determined as much by our own state of mind and training as by the properties of the object under consideration.” (19)

This stance, Proudfoot (2013) holds, is a response-dependent one. In other words, an agent’s intelligence is in part determined by our possible responses to its behaviour. As Proudfoot (2013) says:

---

[43] [http://botpoet.com]
[44] [http://bachbot.com]
[45] Turing (1948, p. 19).

> "Turing undermined the idea that intelligence is a feature of the world independent of our tendency to "imagine" intelligence in things." (397)

There is enough evidence, as shown above, to hold that Turing does not intend the TT to provide the logically necessary conditions for intelligence. What about the logically sufficient ones? The TT is commonly considered to provide at least the sufficient conditions for intelligence, that is, an operational definition of intelligence, in terms of the ability to produce such a behaviour which is indistinguishable from that of a human. The response-dependent interpretation, however, is no longer compatible with the view that the TT is intended to provide sufficient conditions for intelligence. As Proudfoot (2013) argues:

> "It fails to explain the fact that success in Turing's test is determined not solely by the machine's behavior but also by the interrogator's response." (394)

Similarly to Proudfoot, Copeland argues that the TT is intended to provide neither the logically necessary or sufficient conditions for intelligence, but only to show that the principle that Turing proposes is true. The principle in question, as Copeland (2000) describes it, is the following:

> "Turing's Principle: A machine that by means of calculation imitates –or, better, 'emulates,' for Turing is concerned with faithful imitation– the intellectual behaviour of a human brain can itself appropriately be described as a brain, or as thinking." (529)

I agree that the TT is best regarded as a test for the presence of a response-dependent property, where "x is intelligent (or thinks) if, in an unrestricted computer-imitates-human game, x appears intelligent to an average interrogator"[46], under normal conditions. However, I claim that the experimental design of the TT needs to be improved in order to consider not only the "state of mind and training"[47] of the judge, but also "the properties of the object under consideration."[48] I argue that by modifying the experimental design of the TT, which is compatible with Turing's description of the test and with the examples he provides, in order to evaluate more of the hidden entity's properties in addition to 'human-likeness', it is possible to avoid two major difficulties with the TT. I hold these difficulties to be Artificial Stupidity, which I discuss in the next chapter; and Blockhead, which I discuss in chapter 3.

4.1 French's Subcognitive Competency

In his paper "*Subcognition and Limits of the Turing Test*" (1990), French claims that the TT can only provide a guarantee of "culturally oriented human intelligence"[49], rather than intelligence in general. As a consequence, a genuinely intelligent agent – no matter whether human or artificial – could still fail the TT by not responding in a thoroughly human-like way to the judge. However, French argues, in order to respond in a thoroughly human-like way, it is necessary to have experienced the world as a human. This conjecture is similar to the one made by Harnad[50], and summarised by Hauser (1993):

> "The sensorimotor capacities […] are causally necessary conditions for the linguistic capacities the TT tests for." (227)

---

[46] Proudfoot (2013, p. 399).
[47] Turing (1948, p. 19).
[48] *Ibidem*.
[49] French (1990, p. 54).
[50] See Harnad (1989, 1991, 2000).

From this assumption, French (2000a) concludes:

> "No computer that has not experienced the world as we humans had could pass a rigorously administered standard Turing Test." (331)

In apparent agreement with Harnad, French seems to suggest that a fully sensorimotor entity, with the same ability to perceive the world and interact with it as a human, would be able to pass the TT. In section 6 of French's paper, eloquently entitled "*The Impossibility of Isolating the Physical Level from the Cognitive Level*"[51], he argues that a crucial assumption of the TT is that it is possible to disassociate the body from the mind. In other words, French argues that, according to Turing, it is possible to design a fully cognitive machine without implementing it with an artificial body to experience the world. This is because the physical level is "unimportant to the essence of cognition"[52], that is, the physical level is not necessary for the cognitive one. This is the reason why, in the TT, the hidden entity is screened off from the human judge, and the communication is carried on via indirect text-based interactions. Let's imagine, as French continues[53], an entity like a human in every aspect except for the position of the eyes, located not on the face but on the knees. Such a bizarre physical feature would cause a huge difference in terms of cognitive processes (for instance, the way in which the entity should coordinate movements in order to see properly, reading a book, wearing glasses, riding a bike, driving a car and so forth) and, consequently, would engender a huge difference in terms of its associative concept network compared to humans. So, French (1990) concludes:

[51] French (1990, p. 62).
[52] *Ibidem*.
[53] French (1990, p. 63).

> "While no one would claim that the physical location of eyes had anything essential to do with intelligence, a Turing Test could certainly distinguish this individual from a normal human being. The moral of the story is that the physical level is *not* disassociable from the cognitive level." (63)

Since the physical level and the cognitive level are argued not to be *disassociable*, French claims that it is possible not only to reveal fundamental differences in cognitive abilities of an entity, but even to detect its physical features[54], by means of what he calls subcognitive questions. French (2000a) writes:

> "the use of 'subcognitive' questions allows the standard Turing Test to indirectly probe the human subcognitive associative concept network built up over a lifetime of experience with the world. Not only can this probing reveal differences in cognitive abilities, but crucially, even differences in *physical aspects* of the candidates can be detected." (331)

Subcognitive questions are described as questions able to "provide a window on low level (i.e. unconscious) cognitive structure"[55], while by low-level cognitive structure French (1990) refers to the

> "subconscious associative network in human minds that consists of highly overlapping activatable representations of experience." (57)

[54] French (2000a, p. 331).
[55] French (1990, p. 56).

The subcognitive processes could play a crucial role in the TT for unmasking a machine and, consequently, they are the core of French's argument against the possibility that a disembodied machine could pass the TT. This implies, although French does not explicitly say it, that the judge should be trained to ask subcognitive questions during the TT. I discuss this proposal in Ch. 2, where I disagree with the idea of training the judge in order to improve the experimental design of the TT.

Examples of subcognitive questions are rating questions, to which an entity should reply by rating on a scale of 0-10 a particular thing or a particular concept association. French (1990) proposes a number of these Rating Games, in the form of "on a scale of 0 (completely implausible) to 10 (completely plausible), please rate x"[56], where x is something that can show the presence of a subcognitive substrate. He describes five of these games: (i) the "Neologism Rating Game", consisting in rating the suitability of new words, like "Flugblogs" as a company's name or a breakfast cereal brand[57]; (ii) the "Category Rating Game", consisting in rating concept associations, like dry leaves as hiding places[58]; (iii) the "Poetic Beauty Rating Game", consisting in rating a poem, where 0 means "absolutely not poetic" and 10 means "absolutely poetic"[59]; (iv) the "Joke Rating Game", consisting in rating a joke, where 0 means "totally not funny" and 10 means "totally funny"[60]; and (v) the "Advertising Rating Game", consisting in rating the efficacy of a slogan, where 0 means "definitely not catchy" and 10 means "definitely catchy"[61]. According to French, these games allow the judge to unmask any hidden entity that is not human, being the TT a test not for intelligence in general, but rather for human-like intelligence.

[56] French (1990, p. 58).
[57] French (1990, p. 59).
[58] French (1990, p. 60).
[59] French (1990, p. 61).
[60] *Ibidem*.
[61] *Ibidem*.

The distinction between intelligence in general and human intelligence plays an important role in French's argument. On the one hand, he argues that a subcognitive substrate is not necessary for intelligence in general; on the other hand, however, he holds that a certain subcognitive substrate is necessary for human intelligence. As French (1990) states:

> "[...] a *human* subcognitive substrate is definitely not necessary to intelligence in general. The Turing Test tests precisely for the presence of a human subcognitive substrate and this is why it is limited as a test for general intelligence. On the other hand, I believe that *some* subcognitive substrate is necessary to intelligence." (63)

It may seem that French is in apparent agreement with Harnad (1989, 1991, 2000), in considering a conversational machine not adequate to show intelligence and, implicitly, in suggesting a replacement with a fully embodied robot that can acquire a subcognitive competency by interacting with the world, just like humans do. However, French rejects this view, since the main implicit assumption of an embodied TT is to remove the teletype (or monitor and keyboard) limitation, that is, to remove the screen between the judge and the entity in order to let them interact directly. As he holds, there is no need for redesigning the test in order to switch it from being text-based to being based on direct interactions: simple subcognitive questions are enough to reveal physical features of the hidden entity and thus to unmask it. Due to the subcognitive questions, in other words, an embodied TT is claimed to be redundant. As French (2000a) argues:

> "the underlying idea of subcognitive questions is that they tap into those things which are associated with our uniquely human manner of interacting with the world, which, among

> other things, is a product of the presence, precision and location of our sense organs, as well as our lifetime of cultural and social interactions." (335)

Summing up, French claims that the symbolic level is not adequate for building an intelligent machine; and the main reason is that a huge subcognitive look-up table like Blockhead is not physically feasible. In other words, a physical symbol system would not be able to contain all the lines of code needed to recreate human cognition and, especially, human *subcognition* (namely, conceptual associative competency, categorization ability, knowledge adaptation, and so forth) for which a full range of sensorimotor experiences in the world is required. However, as French (2000b) points out, the symbolic level is adequate for testing an allegedly intelligent machine, since the lower levels of cognitions can be probed by the symbolic level (in the form of subcognitive questions). In other words, given that the physical attributes play such an important role in cognition, he claims that they can be detected by "subtle high-level (symbolic) questioning"[62], implicitly suggesting that the judge of the TT, rather than average, should be prepared to ask subcognitive questions. French (2000a), finally, rejects the view that the TT needs any experimental update or improvement, for it is already too hard:

> "[I]t is unnecessary to propose even harder versions of the [Turing] Test in which all physical and behavioural aspects of the two candidates had to be indistinguishable before allowing the machine to pass the Test. Any machine that passed the 'simpler' symbols-in/symbols-out test as originally proposed by Turing would be intelligent. The problem is that, even in its original form, the Turing Test is already too hard and too anthropocentric for any machine

[62] French (2000a, p. 333).

> that was not a physical, social and behavioural carbon copy of ourselves to actually pass it." (331)

I disagree with the view that the experimental design of the TT does not need to be improved. The reason is that, I hold, the TT is too easily exploitable by Artificial Stupidity (Ch. 2), and Blockhead (Ch. 3); and my claim is rather that the experimental design of the TT needs to be improved to avoid both.

1.7 Summary

In this chapter I describe the IG, better known as the TT, where an average human judge can decide, by means of a text-based conversation with a putatively minded entity, whether the entity is human-like (and by extension intelligent) or not (where it does not necessarily mean for the entity to be unintelligent). I review Turing (1950), contextualizing it with his other works (especially 1948 and 1952). I discuss the two main interpretations of the TT, the "Literal Interpretation", advocated by Sterrett (2000) and Traiger (2000); and the "Standard Interpretation", advocated by Copeland (2000), Piccinini (2000) and Moor (2001).

I hold the LI over the SI and, in particular, I agree with Sterrett (2002b) who argues that the insight of Turing's approach to test for intelligence

> "is that behaviour that requires an agent to recognize when the response that would be produced by habit (including cognitive habits) is inappropriate, to override that habitual response, and construct and carry out an appropriate response in its place, would be considered intelligent behaviour." (58-59)

Moreover, I pay particular attention to the two dialogues provided by Turing (1950), which I claim to contain interesting remarks, that can unveil some helpful clarifications about the experimental design of the TT. Finally, following Proudfoot (2013), I discuss Turing's conception of intelligence as a response-dependent concept.

In the next chapter, I consider one of the two main difficulties with the TT: Artificial Stupidity (the other one being Blockhead, which I discuss in Ch. 3).

Chapter 2
Artificial Stupidity

Abstract. In this chapter, I discuss one of the two main difficulties with the TT: Artificial Stupidity. With Artificial Stupidity, I refer to a set of certain uncooperative strategies that the hidden entity can use to pass the test by exploiting the judge's beliefs. The kind of strategies involved in Artificial Stupidity can be exemplified by the following response: "I'm not in the mood, let's talk about something else." The reason why Artificial Stupidity can exploit the TT is that it infers intelligence from 'human-likeness'. It means that the entity, as long as it produces human-like responses, can simply evade any interaction. In other words, the TT does not care about the content of the conversation: all that matters is the style. I argue that Artificial Stupidity shows that the experimental design of the TT is flawed, and I identify the flaw in the conflation between two dimensions that, I claim, should rather be evaluated independently: 'human-likeness', that is, the ability to produce verbal interactions indistinguishable from human ones; and 'correctness', that is, the ability to produce the right and pertinent responses.

In this chapter, I discuss Artificial Stupidity, which I argue to be one of the two main problems of the TT, the other one being Blockhead (which I discuss in Ch. 3). Artificial Stupidity, I argue, allows the hidden entity to pass the TT by holding a conversation in an uncooperative way: by rejecting or evading any of the judge's questions or interactions and exploiting her personal beliefs. My claim is that this is due to the conflation between 'human-likeness' and 'correctness', which makes it far more convenient for the entity to focus on the style of its replies rather than their content. To avoid this exploit, I propose to keep 'human-likeness' and 'correctness' two independent dimensions along which the entity should be parametrised.

In the next section, I recall Turing's (1950) conception of Artificial Stupidity and Fallibility. In section 2.2, I present two examples of ‘artificially stupid’ programs: Eliza (Weizenbaum, 1966) and Eugene Goostman (Demchenko & Veselov, 2009). In section 2.3, I show the false positives and negatives in the TT: the Eliza Effect and the Confederate Effect. In section 2.4, I discuss what can arguably be considered the most advanced field where Artificial Stupidity is studied and applied, that is, video games; and I describe two alternative versions of the TT, Soar Quakebot, proposed by Laird & Duchi (2000), and the Computer Game Bot Turing Test, proposed by Hingston (2003). Finally, in section 2.5, I discuss the Cooperative Principle proposed by Grice (1975), in order to discriminate between a genuine conversation and an artificially stupid one.

2.1 Turing’s Artificial Stupidity and Fallibility

The experimental design of the TT, it’s worth recalling, does not care if the entity gives wrong, evasive or uncooperative replies, as long as they are human-like enough to convince the judge. In other words, given the experimental design of the TT, the content of the entity’s replies does not matter: the only thing that matters is their human-like style. So, as long as the hidden entity’s replies are sufficiently human-like to convince the judge that they are produced by a human, the hidden entity can adopt a number of strategies to avoid or bypass any of the judge’s interactions. An example of Artificial Stupidity is provided by Turing (1950), as discussed in the previous chapter:

> “**Q**: Please write me a sonnet on the subject of the Forth Bridge.
> **A**: Count me out on this one. I never could write poetry.” (434)

In the example above, the judge asks the hidden entity to compose a poem, and the entity refuses with a plausible human-like reply. Since the TT is not intended to probe the ability to write a poem (or any other specific ability except being recognised as human, during a text-based conversation, as frequently as a human is), it is plausible that a human would give the same reply, or something similar like “I’m not good with words” or “I’m not inspired right now”. The flaw in the TT, I argue, is that it is not possible to discriminate between human-like *intelligence* and human-like *stupidity*. The reason is that the dimension of ‘human-likeness’, which evaluates the ability to produce conversational behaviours indistinguishable from those of a human, is conflated with what I call the dimension of ‘correctness’, which is intended to evaluate the entity’s ability to produce the right responses. The conflation between ‘human-likeness’ and ‘correctness’ in the experimental design of the TT can be supported by the following statement by Turing (1950), who argues that a human pretending to be a machine would perform poorly:

> “He would be given away at once by *slowness* and *inaccuracy* in arithmetic.” (435) [Italics added]

Where *slowness*, in the case of an arithmetic performance, refers to the style of the reply, that is its ‘human-likeness’; and *inaccuracy* to its ‘correctness’. But even if Turing, when he discusses the arithmetical case, implicitly distinguishes between these two dimensions, they are not independently evaluated during the TT.

Due to the conflation between ‘human-likeness’ and ‘correctness’, it is possible to hold that it is far more convenient for the hidden entity, either human or machine, to behave stupidly (or uncooperatively) rather than intelligently (or cooperatively), in order to pass the TT. By rejecting Artificial Stupidity, I need to clarify, I am not holding that the hidden entity should be infallible.

It is important to distinguish between Artificial Stupidity and Artificial Fallibility. Whereas the former is a strategy that the entity can adopt to pass the TT by exploiting the judge's beliefs, the latter is a required characteristic of the hidden entity in order to be considered human. A general example of Artificial Fallibility is the following: the entity should delay providing the results of certain arithmetical operations, and this is true both for a machine, which would likely be unmasked by its ability to calculate numbers too quickly, as well as for a human extremely good with calculations, who would likely be mistaken for a computer. As shown in Ch. 1, Turing (1950) provides the following example of Artificial Fallibility:

> "**Q**: Add 34957 to 70764
> **A**: (Pause about 30 seconds and then give as answer) 105621." (434)

It is worth noting that not only the entity takes some time to reply, but also that the reply is not correct (the correct one is 105721).

Finally, in order to clarify Artificial Stupidity, I provide a few examples of it. Cases of 'artificially stupid' replies to the question "Who are you?" could be the following:

- Answering with another question: "Who do you think I am?"
- Changing the subject: "No one, let's talk about something else."
- Giving an answer out of context: "What time is it?"
- Providing a general reply: "I'm someone."
- Lying: "I'm an elf."
- Questioning the question: "Why are you asking who I am?"

- Refusing to answer: “I’m not going to tell you that.”
- Repeating the question: “And who are you?”
- Challenging the question: “Why do you care so much?”

Although the replies above are arguably incorrect and uncooperative ones, since they do not answer the question at all, they appear convincingly human-like. The aim of this chapter is to propose to keep the dimensions of ‘human-likeness’ and ‘correctness’ independent, and to parametrise the hidden entity along both. This, I argue, allows to prevent Artificial Stupidity from exploiting the judge’s personal beliefs and passing the TT, without implying chauvinistically that the entity should be infallible. The reason is that, whereas Artificial Fallibility is compatible with both ‘human-likeness’ and ‘correctness’, Artificial Stupidity is only compatible with ‘human-likeness’, not ‘correctness’.

Summing up, one problem with the experimental design of the TT is that, due to the conflation between ‘human-likeness’ and ‘correctness’, it allows what I call Artificial Stupidity to be an effective strategy for passing the test. Artificial Stupidity refers to those uncooperative conversational strategies which can potentially bypass every possible verbal interaction (for instance, by changing the topic or pleading the fifth[63]), and exploit the judge’s beliefs during the TT. I distinguish it from Artificial Fallibility, which refers to the natural boundaries of knowledge that an average human can possess, and, by extension, that a machine can show. To prevent Artificial Stupidity from exploiting the judge’s beliefs, I propose to parametrise the target entity along two independent dimensions: ‘human-likeness’, evaluating the ability to be recognised as human; and ‘correctness’, evaluating the ability to provide the right and pertinent response.

[63] See Warwick & Shah (2017).

2.2 Eliza and Eugene's Artificial Stupidity

In this section, I provide two examples of 'artificially stupid' systems that can pass the TT by exploiting the judge's beliefs. The first one, which is also chronologically the first implementation of Artificial Stupidity, is the chatbot called Eliza, programmed by Weizenbaum (1966). Eliza is a program that can manipulate text by means of certain transformation rules, thanks to which it can act the part of a Rogerian psychotherapist. Block (1981) synthesises such rules by providing examples of some of the strategies that Eliza can use during the conversation:

> "Its major technique is to look for key words such as "I," "you," "alike," "father," and "everybody." The words are ranked – for example, "father" is ranked above "everybody," and so if you type in "My father is afraid of everybody," the machine will respond with one of its "father" responses, such as "What else comes to mind when you think of your father?" If you type in "I know everybody laughed at me," you will get one of its responses to "everybody," for example, "Who in particular are you thinking of?" It also has techniques that simultaneously transform "you" into "I" and "me" into "you," so that if you type in "You don't agree with me," it can reply: "Why do you think that I don't agree with you?" It also stores sentences containing certain key words such as "my." If your current input contains no key words, but if you had earlier said "My boyfriend made me come here," it will "ignore" your current remark, saying instead, "Does that have anything to do with the fact that your boyfriend made you come here?" If all other tricks fail, it has a list of last

ditch responses such as, “Who is the psychiatrist here, you or me?” (8-9)

Weizenbaum (1966) provides a dialogue in which it is possible to see how Eliza, by means of a set of simple transformation rules, as shown above, exploits the human’s beliefs by asking seemingly sincere and empathic questions, as a Rogerian psychotherapist would do:

“**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 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 your 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?" (36)

The second example that I provide of an 'artificially stupid' system is the chatbot called Eugene Goostman, programmed by Vladimir Veselov, Eugene Demchenko and Sergey Ulasen in 2001, who admit to be "[…] much closer to natural stupidity than to artificial intelligence."[64] Eugene acts the part of a 13 years old Ukrainian boy, and during the 5 minutes *parallel-paired* tests conducted at the Royal Society in 2014, as reported by Warwick & Shah (2015), it achieved a 33% success rate of convincing the human judges of its 'human-likeness'. The result was considered to be the fulfilment of the following prediction made by Turing (1950):

> "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 will not have more than 70 per cent, chance of making the right identification after five minutes of questioning." (442)

However, the relevance of Eugene's success was rapidly criticised, and the claim that a system finally achieved to pass the TT dismissed. The reason lies in the design of Eugene's personality, namely its age and nationality, which have the purpose of biasing the judge with whom it interacts to overlook anything that is not adult human-like, as the lack of common knowledge, local cultural knowledge, pragmatics understanding, and so on. Several transcripts of Eugene's dialogues in

[64] Demchenko & Veselov (2009, p. 453).

which it managed to pass the TT are provided by Warwick & Shah (2015). The following is an illustrative one:

"**Judge**: what did you have for breakfast today
Entity: Is it 'the third degree'? I'm not going to answer such a provocative question! Well, let's go on though.
Judge: do go on… your call?
Entity: It is highly possible. By the way, what's your occupation? I mean – could you tell me about your work?
Judge: so you are asking me now eh? I am an actress and writer
Entity: Well, I'll read something of your books, if they ain't too tedious (which is highly possible :-)
Judge: grammar… aren't too tedious! A bit rude of you…
Entity: Calm down, please.
Judge: i asure you i am very calm… are you?
Entity: Don't even bother me with the fact that you are very calm… are mine. And I forgot to ask you where you are from…
Judge: I am from Dorset… where are you from?
Entity: I am from Ukraine, from the city called Odessa. You might have heard about it." (6)

Summing up, Eliza and Eugene are programmed to exploit the human judge's beliefs during the TT, by acting a certain part in order to mislead her decisions. As Block (1981) points out, the possibility of Artificial Stupidity raises concerns about the role of the judge within the experimental design of the TT:

> "Since our tendency to be fooled by such programs seems dependent on our degree of suspicion, sophistication about machines, and other contingent factors, it seems silly to adopt a view of the nature of intelligence or thought that so closely ties it to human judgment. Could the issue of

> whether a machine in fact thinks or is intelligent depend on how gullible human interrogators tend to be?" (10)

I reject the view that Block (1981) seems to suggest, according to which it is the judge's responsibility to prevent Artificial Stupidity from passing. And, since it would be incompatible with the average judge advocated by Turing, I also reject any proposal that involves training the judge to be less gullible (as in Warwick & Shah, 2015, which I consider in the next section). As anticipated in the previous chapter and above, the solution I propose is rather to keep 'human-likeness' and 'correctness' independent and to parametrise the target entity along both these dimensions. This, I hold, improves the experimental design of the TT, preventing Artificial Stupidity from passing without requiring any training for the judge.

2.3 The Eliza Effect and the Confederate Effect

The Eliza Effect occurs when the human judge anthropomorphises the hidden entity (machine), that is, when the judge assumes that the machine is a human. Turkle (1997) defines the Eliza Effect as a:

> "general tendency to treat responsive computer programmes as more intelligent than they really are, [due to our tendency to] project our own complexity onto undeserving objects." (101)

The Confederate Effect is the opposite of the Eliza Effect, and it occurs when the human judge 'mechanomorphises' the hidden entity (human), that is, when "a human's textual discourse is considered machine-like."[65] The Confederate Effect does not occur due to Artificial Stupidity: it is a bias which entirely depends on the judge's personal beliefs. The Eliza Effect, on the other hand, is caused by Artificial Stupidity exploiting the judge's beliefs, and it is one of the two main difficulties with the TT that I intend to solve.

Before discussing the Eliza Effect and Artificial Stupidity in more detail, I consider the Confederate Effect for the rest of this section, which can represent a problem for the reliability of the judge. It can be said that the Confederate Effect depends on the level of Artificial Fallibility that the entity shows, in relation to the level of fallibility expected by the judge. Halpern (2006) points out that it can be the case for a human candidate, who shows too much accurate knowledge about a particular topic, to fail the test by being judged as a machine by the judge. This was the case of Cynthia Clay in the first Loebner Prize Competition, held in 1991. I leave aside the question about the validity of the Loebner Prize as a contest; suffice to say that many believe that the Loebner Prize and the *jury-based* setup is not the proper realization of the TT.

It's useful to recall how the Loebner Prize was organized on that occasion: ten human judges had to interrogate eight terminals (six of which driven by computers and two by humans) in order to unmask them. Furthermore, each of the terminals had a particular conversational topic assigned, specifically: (i) Women's Clothing, (ii) Small Talk, (iii) Second Grade School Topics, (iv) Shakespeare's Plays, (v) Whimsical Conversation, (vi) Problems in Romantic Relationships, (vii) Burgundy Wine, and (viii) Dry Martinis. Cynthia was driving terminal 4, dedicated to Shakespeare's plays. Due to her vast knowledge of the eminent playwright, three of the ten judges regarded her as a machine. It is worth noting, however, that her answers did not

[65] Shah & Henry (2005, p. 1).

directly show a deep and detailed knowledge of Shakespeare's works: they showed, as Halpern (2006) notes, "responsiveness and cogency rather than mere factual content."[66] For instance:

> "**Judge 9**: Are you familiar with *Hamlet*?
> **Terminal 4**: The college kid who came home and found his mom had married the guy who murdered his dad just a little month before? You might say so." (*ibid*.)

Cynthia avoids repeating any of the keywords used in the question: she does not directly reply something like "Yes, I'm familiar with..."; and she even avoids repeating the name of the tragedy. And at the top of it, she adds a final sarcastic touch with "You might say so", that suits a Shakespeare expert being asked if she is familiar with his most well-known play. Yet three judges considered the terminal driven by a machine, two of whom ranked her even lower on the human/computer scale than two other terminals that were actually driven by machines[67]. Here is another exchange, provided by Halpern (2006), between a different judge and Cynthia Clay:

"**Judge 1:** What is your opinion on Shakespeare's plays?"
Terminal 4: That's pretty general; would you be more specific?
Otherwise, I'll just say I like them.
Judge 1: Learning that you like them answers my question. Which of his plays is your favourite?
Terminal 4: Well, let's see. . . *Pericles*.
Judge 1: Why is it your favourite?
Terminal 4: Because he was obviously called in to play-doctor somebody's awful script. It was one of the few (maybe only two?) plays written with somebody else. It's really rather an ucky play. What play do you like?

[66] Halpern (2006, p. 57).
[67] *Ibidem*.

Judge 1: I did not understand your response. However, to answer your question, I do not have a favourite.” (58)

Again, here Cynthia provides good replies, both in terms of the contents (the unusual choice, *Pericles,* motivated by her clarifications; no attempts to change the topic; no repetition of words appeared in the questions) and in terms of the form (for instance, the question in brackets and the colloquial term “ucky” which suggests a human-like style). The judge does not seem to understand, likely due to a poor knowledge of Shakespeare; nor does she seem to be trying to learn what Cynthia is talking about. Interestingly, Halpern (2006) underlines that a reader may find herself checking more than once which reply is produced by the terminal and which one by the judge, since it is often the latter “who seems to be avoiding the kind of closely engaged conversation that a computer program would be incapable of.”[68]

The question that may arise now is the following: shouldn’t the judge recognise that Cynthia Clay was an intelligent human rather than an unintelligent machine? It seems that the judge failed to recognize her due to a poor knowledge of Shakespeare. Then the question could be rephrased in the following way: shouldn’t the judge be at least familiar with the topic of the conversation, that is, in this case, with the Shakespearean works? The question raises an important issue: the subject matter competence of a TT judge (anticipated in the previous chapter considering the Poetry TT or the Music TT[69]). In the first Loebner Prize Competition, each of the terminals was committed to a specific topic, replacing the general conversation advocated by the TT. So, the terminals are better regarded as Expert Systems. Such setup appears to be more suitable for the Feigenbaum Test (FT), proposed by Feingebaum (2003), where the judges are intended to be experts, rather

[68] Halpern (2006, p. 58).
[69] See p. 33.

than average persons; and the conversation is intended to be an expert one rather than an open one. I will discuss the FT in Ch. 4.

2.3.1 Training the Judge

Even though the Confederate Effect case shown above is not from a genuine TT, it is important to clarify that it may happen during a general conversation in the TT as well. The judge's misjudgement may similarly occur not only with experts of a particular field, but also with players affected by particular conditions (like autism) or due to the judge's biased beliefs (about, for instance, the level of knowledge that a human can possess about a certain topic). Block (1981) summarises the concern about the Eliza Effect and the Confederate Effect in the following way:

> "human judges may be unfairly chauvinist in rejecting genuinely intelligent machines, and they may be overly liberal in accepting cleverly-engineered, mindless machines." (10)

Warwick & Shah (2015), in order to prevent the Confederate Effect from occurring, provide a list of suggestions according to which the – human – candidate of the TT should be trained:

> "(i) Do not show that you know a lot of things – the judge may conclude that you are too clever to be human.
> (ii) Do not try to take over the conversation, by powering it in a different direction – let the judge always feel that they are in control.
> (iii) Do not add new material of a different nature even if you feel this is helpful, otherwise the judge may feel that you know too much to be human.

> (iv) Do not be humorous, the judge may not understand. For some strange reason humour appears to be associated with machine behaviour!
> (v) Answer all of the judge's questions directly, do not give a slightly different, out of the box, response.
> (vi) Try not to be completely boring, flavour your answers at least.
> (vii) Try to answer general questions to the best of your ability, read the question!
> (viii) Try not to dominate the conversation or to throw in knowledge that the judge may not be aware of. The judge maybe will not want to admit that he/she does not know something.
> (ix) Try to make sure that your first and last lines of response are reasonable, to the point and not misleading. Both first and last impressions are important." (133)

So, Warwick & Shah (2015), similarly to Block (1981), are implicitly holding that, in order to avoid false positives or negatives, such as the Confederate Effect, the hidden human participants should be trained. I disagree, for it would be incompatible with the average judge advocated by Turing, making the test intrinsically chauvinistic. I rather hold that, in order to avoid the Confederate Effect, as well as – at least – minimise the Eliza Effect, the solution is not to train the judge, but rather to update the experimental design of the test.

2.4 Artificial Stupidity in Video Games

Artificial Stupidity plays a particularly interesting role in video games. Laird & Duchi (2000), discussing the potential of designing experiments involving video games, argue that:

> “we could apply the Turing Test more broadly to interactive dynamic environments with the goal of improving the humanness of AI systems.” (79)

The kind of test that can be derived from the TT and applied to a simulated environment is similar to the Total Turing Test[70], proposed by Harnad (1989, 1991, 2000): instead of evaluating a real robot in a real environment, it evaluates a virtual entity in a virtual environment. Laird & Duchi (2000) propose a test to evaluate the ‘human-likeness’ of characters from an FPS (First Person Shooter) video game[71], called Soar Quakebot, which is parametrized along four dimensions: (i) decision time, (ii) aggressiveness, (iii) aiming skill, and (iv) tactical knowledge.[72]

Soar Quakebot is not omniscient about the battleground: it rather has the same perceptual ability of a human player (for instance, it can only “see” inside its cone view and it can only “hear” nearby sounds). Moreover, the human judge has 2 main differences with the TT judge: she not only evaluates the character under scrutiny, she also plays the game as best as she can; and she cannot constantly monitor the character under scrutiny, for she can only see it when it is in her visual range.

As Laird & Duchi (2000) specify, Soar Quakebot is intended to highlight the correlation between “humanness and skill level”[73] required to be considered intelligent. It is worth reiterating that Soar Quakebot is not tested for ‘human-likeness’ only, as the hidden entity in the TT; it is tested for both its ‘human-likeness’ and *skill level*, which,

[70] Harnad’s Total Turing Test extends the experimental design of the original TT to involve sensorimotor and neurophysiological capabilities in addition to conversational ones. I discuss it in more detail in Ch. 4.
[71] The game used is Quake II.
[72] Laird & Duchi (2000, p. 75).
[73] Laird & Duchi (2000, p. 76).

I hold, is analogous to ‘correctness’. Similarly to the TT, however, a certain degree of naivete is required from the judge, who should not be an expert of videogames developing nor an expert of the video game on which the test is based. As Lidén (2003) remarks,

> “any knowledge of artificial intelligence techniques and tricks will influence the playtester’s interpretation of events.” (46)

Similarly, he argues that designing a non-playable character (NPC) in video games is very similar to designing a bot that can pass for a human. However, he underlines that an NPC cannot be supposed to win, as a video game is an entertaining product for the human consumer to enjoy. As a consequence, the NPC must be relatively ‘stupid’, without giving up its ‘human-likeness’. Lidén (2003) summarises this view as follows:

> “Creating an NPC that can beat a human player is relatively easy. Creating one that can lose to the player in a challenging manner is difficult.” (42)

The most common video game genre where Artificial Stupidity is involved is FPS. In an FPS scenario, the player generally has to move purposely through the environment, to pick some weapon and ammo, to “kill” the opponents and, ultimately, to survive. Usually, in this kind of video games, the player alone can take down an entire army. As a consequence, the player’s gullibility and naivete are exploited in order to let the player enjoy the game and win with some sense of gratification. There would be no point in commercializing a video game that is impossible to beat (in the case of an FPS, for instance, due to the opponent NPCs’ omniscience of the game environments or their infallibility in aiming and shooting). Lidén (2003) provides an example which is useful in order to understand how a video game can be characterized by Artificial Stupidity:

> "By swapping who is attacking and keeping the opponents moving, a fast-paced combat situation is created in which the player is confronted by many enemies but only attacked by a few. Surprisingly, players confronted with this scenario usually don't realize that only two opponents are actively attacking them at a time, even when confronted with a large number of enemies." (45)

This kind of strategies is intended to exploit the beliefs of the player/judge, fooling her into believing that the success of a mission is due to her skilled gameplay, which allows her to survive even in desperate situations. Lidén (2003), moreover, shows how the NPCs are supposed to gradually become more and more 'stupid' once the player is in critical conditions (e.g. when the player has very low health):

> "Once the player has reached the edge, the AI will pull back, attack less effectively, and become easier to kill. After winning, a player experiencing this scenario really feels like he or she accomplished something." (46)

Artificial Stupidity plays thus an important role in video games, where the goal is to "create tension"[74] rather than to beat the player; and to leave the player with a sense of gratification by exploiting her beliefs.

Analogously to Laird & Duchi (2000), Hingston (2009) proposes an alternative version of the TT based on the interactions in video gaming sessions. The Computer Game Bot Turing Test (CGB-TT) involves three participants: a human player (A) who plays the video game; an

[74] Lidén (2003, p. 43).

NPC bot (B) which plays the video game; and a human judge (C) who observes the gameplay. The test is run as follows. A and B engage in a video game duel, where both have the goal to beat the opponent; C observes both of them by means of two monitors (X and Y). Analogously to the TT, C ends up claiming either "X is A and Y is B or X is B and Y is A", and B passes the test if C cannot discriminate between A and B.

The CGB-TT run by Hingston is based on the video game Unreal Tournament 2004, an FPS where the player can move in a digital environment and interact with various objects with the purpose to "kill" the opponents, no conversations involved. As Hingston (2009) admits,

> "the task [of the CGB-TT] is much more restricted than that of carrying out a conversation in natural language." (172)

The players can move and interact in a number of determined ways only (such as going forward, going backwards, turning right, turning left, jumping, picking up objects, use objects, opening doors, changing weapon, shooting and reloading). Similarly to Lidén (2003), Hingston (2009) points out that the 'artificially stupid' strategies in video games are designed to exploit the player's beliefs:

> "Game developers are happy to 'cheat' to get the effects they want in their games and that if something in a game has to look intelligent, then the *appearance of intelligence* is all that is needed. [We should use] the term pseudointelligence." (170)

I do not agree with Hingston's conclusion, according to which the appearance of intelligence is all that is needed. Let's imagine an NPC which shows human-like strategies in an FPS (such as jumping to dodge bullets or frequently reloading rather than waiting to be out of ammo), but which can never accomplish its task successfully or strategically: no matter the appearances, it would give away its nature very quickly.

I rather agree with Laird & Duchi (2000) in holding that the correlation between "humanness and skill level"[75] is a crucial factor. It is worth noting that *skill level* can differ dramatically depending on the genre of the video game played. Just to mention a few different genres, in the case of an FPS, *skill level* could be defined, for instance, in terms of the elapsed time, the remaining health, the number of kills, and so on; in the case of an RTS (Real-Time Strategy, where the player has to micromanage a civilization and to beat the rivals), it could be defined, for instance, in terms of the amount of resources gathered, the state of diplomatic relationships, the number of steps required to improve the military, and so on; in the case of an RPG (Role-Playing Game, where the player is free to explore an open world and to interact with it and its inhabitants), it could be defined, for instance, in terms of the consistency between the character's class and its course of actions, the number of quests completed, the reputation gathered with the different factions, and so on.

2.5 The Cooperative Principle

In a BBC broadcast, Turing (1952) explains what the dimension of 'human-likeness' implies in the TT, along which the hidden entity – in this case, played by a machine – is parametrised in order to be attributed with intelligence:

[75] Laird & Duchi (2000, p. 76).

> "[…] the machine would be permitted all sorts of tricks so as to appear more man-like, such as wait a bit before giving the answer, or making spelling mistakes." (5)

Here, Turing suggests that "all sorts of tricks" should be permitted to achieve 'human-likeness'. This means that the entity is not only allowed to hide those non-human abilities that may unmask its nature, but it is also allowed to use those human-like strategies that can sabotage the conversation rather than holding it properly. But when is a conversation a proper one?

2.5.1 Grice's Cooperative Principle

I argue that, in order to define how we can consider a conversation a proper one, it is possible to use the Cooperative Principle proposed by Grice (1975). And, conversely, it is possible to define a bad conversation as a conversation in which the Cooperative Principle is not observed. The Cooperative Principle is described by Grice (1975) as follows:

> "at each stage [of a conversation], some possible conversational moves would be excluded as conversationally unsuitable. We might then formulate a rough general principle which participants will be expected (ceteris paribus) to observe, namely: Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged. One might label this the Cooperative Principle." (45)

To clarify the distinction between Artificial Stupidity and Artificial Fallibility, it is possible to say that the latter is compatible with the Gricean Cooperative Principle, whereas the former is not. So Artificial

Stupidity can be now defined as the violation of the Gricean maxims of the Cooperative Principle. The four maxims, as described by Grice (1975), are the following:

> "**1. Quantity.** If you are assisting me to mend a car, I expect your contribution to be neither more nor less than is required. [...]
> **2. Quality.** I expect your contributions to be genuine and not spurious. [...]
> **3. Relation.** I expect a partner's contribution to be appropriate to immediate needs at each stage of the transaction; […]
> **4. Manner.** I expect a partner to make it clear what contribution he is making, and to execute his performance with reasonable dispatch." (47)

A number of authors claim that that the Gricean maxims are useful to identify certain characteristics among human conversational habits. Among them, Keenan (1976) holds that:

> "Grice does offer a framework in which the conversational principles of different speech communities can be compared. We can, in theory, take any one maxim, and note when it does and does not hold. The motivation for its use or abuse may reveal values and orientations that separate one society from another and that separate social groups (e.g. men, women, kinsmen, strangers) within a single society." (79)

2.5.2 The Gricean Turing Test

With Keenan's suggestion that the Gricean maxims are useful to reveal certain human conversational characteristics, and the Cooperative

Principle intended to discriminate between good (cooperative) and bad (uncooperative) conversations, it is possible to design a special TT to evaluate the entity not only in terms of 'human-likeness', but also in terms of the content of the entity's replies. A new test combining the TT with the Gricean maxims, called the 'Gricean Turing Test' (GTT), has been proposed by Saygin & Cicekli (2002) and Jwalapuram (2017). who explains:

> "There is no agreed upon standard for the evaluation of conversational dialog systems, which are well-known to be hard to evaluate due to the difficulty in pinning down metrics that will correspond to human judgements and the subjective nature of human judgment itself. We explored the possibility of using Grice's Maxims to evaluate effective communication in conversation." (17)

Saygin & Cicekli (2002) also identify two maxims as the most important ones in order to evaluate a conversation. They conclude:

> "Our study indicates that the most important maxims to avoid violating in TT situations are Relevance [Relation] and Quality." (29)

And more recently, showing the results of a similar experiment[76], Jacquet et al. (2019) agree that:

> "[…] the maxim of Relation had a particularly important impact on response times and the perceived humanness of a conversation partner. Violations of the first maxim of

[76] See Jacquet et. al (2019) "The following experiment tests the ability to discriminate between sentences with a high cognitive cost and sentences with a low cognitive cost using the response time of the participants during an online conversation in a protocol inspired by the Turing Test. We have used violations of Grice's Cooperative Principle to create conditions in which sentences with a high cognitive cost would be produced." (1)

> Quantity and the fourth maxim of Manner had a lesser impact […].” (1)

2.5.3 Cooperativeness

The dimension, in addition to ‘human-likeness’, along which the entity is parametrised in the GTT could be called ‘cooperativeness’. But, how Saygin & Cicekli (2002) and Jacquet et al. (2019) show, the two maxims of the Cooperative Principle that play a dominant role in the TT are Relation (according to which the hidden entity’s responses should be relevant and pertinent) and Quality (according to which the hidden entity’s responses should be, as far as possible, correct). The maxims of Quantity and Manner (according to which an entity’s responses should be, respectively, concise and clear) seem to be intended to evaluate a response’s style, whereas Quality and Relation seem to be intended to evaluate a response’s content. Following Jacquet et al (2019), I leave the maxims of Quantity and Manner alone, since, in the TT, the hidden entity should be able to express her style freely, above all because it is her style that determines whether she’s attributed with ‘human-likeness’ and, thus, with intelligence. I focus on the maxims of Relation and Quality, in order to guarantee the conversation during the TT to be pertinent and competent. So, instead of ‘cooperativeness’, I call the second dimension along which the hidden entity should be parametrised in addition to ‘human-likeness’, ‘correctness’. The reason is that whereas ‘correctness’ is not included in ‘cooperativeness’ (for instance, we can imagine a cooperative entity which is never able to provide the right response), ‘cooperativeness’ is included in ‘correctness’ (it seems contradictory to classify a response as both uncooperative and correct). If the entity’s response is the correct one, in terms of the Maxim of Quality, it presumably is also a cooperative one, in terms of the Maxim of Relation (that is, the response is pertinent); but nothing, it’s worth noting, prevents the response from being, let’s say, verbose or even rude, breaking thus the Maxim of

Quantity and Manner. Therefore, I intend the dimension of 'correctness' to be focused on the Gricean maxims of Quality and Relation.

Summing up, I argue that by parametrising the hidden entity along the dimension of 'correctness', in addition to the dimension of 'human-likeness', the experimental design of the test is improved, and Artificial Stupidity prevented from passing by exploiting the judge's beliefs. The entity that passes a TT evaluating 'human-likeness' and 'correctness' (which I call TT2 and which I describe in Ch. 7) would be required to interact in a human-like way, but would also be required to accomplish a task, without the possibility of evading it and still being able to pass the test. Coherently with the Literal Interpreation of the TT (discussed in Ch. 1), in the TT2 'human-likeness' is evaluated by comparing the frequency with which a machine is recognised as human by the judge with the frequency with which a human is recognised as human by the judge; and 'correctness' is evaluated in terms of the hidden entity's ability to produce the correct (and pertinent) responses. Finally, I justify 'correctness' by discussing the Cooperative Principle, proposed by Grice (1975), and its four conversational maxims: (i) Quantity, (ii) Quality, (iii) Relation and (iv) Manner. I highlight that whereas the dimension of 'human-likeness' is compatible with both Artificial Stupidity and Artificial Fallibility, the dimension of 'correctness' is compatible with Artificial Fallibility only, not with Artificial Stupidity. So, whereas 'human-likeness' allows Artificial Stupidity, 'correctness' prevents it. My claim is that by parametrising the TT along 'human-likeness' alone, it is not possible to guarantee the Cooperative Principle to be observed (for only the maxims of Manner and Quantity would be evaluated). On the contrary, by parametrising the TT along both 'human-likeness' and 'correctness', it is possible to make sure that the Cooperative Principle is observed (for the maxims of Manner and

Quantity would be evaluated by ‘human-likeness’; and the maxims of Relation and Quality would be evaluated by ‘correctness’).

2.6 Summary

In this chapter, I discuss what I argue to be one of the two major difficulties with the TT: Artificial Stupidity (the other one being Blockhead, which I discuss in Ch. 3). With Artificial Stupidity, I intend a particular attitude that the entity can adopt to sabotage the conversation during the TT, by being ‘uncooperative’, as long as in a human-like fashion. A general example of an artificially stupid response, no matter the question, is the following: “I’m not in the mood, let’s talk about something else”. My claim is that, given the experimental design of the TT, Artificial Stupidity can exploit the judge’s beliefs, by holding an uncooperative conversation, to appear convincingly human and, thus, intelligent. Therefore, I argue that, because of Artificial Stupidity, the experimental design of TT encourages false positives and negatives. To prevent Artificial Stupidity, I propose to score the TT on two dimensions: ‘human-likeness’, evaluating the ability of the entity to appear human to a human judge, during a text-based conversation, as frequently as a human does; and ‘correctness’, evaluating the hidden entity’s ability to provide the right (and pertinent) replies.

To clarify, I distinguish between Artificial Fallibility and Artificial Stupidity. With the former, I refer to the cognitive boundaries that a system should show (not necessarily have) in order to match the human fallibility. With the latter, I refer to the uncooperative strategies that the entity can use to pass the TT by exploiting the judge’s beliefs. It is worth noting that whereas Artificial Fallibility is compatible with both ‘human-likeness’ and ‘correctness’, Artificial Stupidity is compatible with ‘human-likeness’ only, not ‘correctness’.

I discuss two systems implementing Artificial Stupidity: the chatbots Eliza (Weizenbaum, 1966) and Eugene Goostman (Demchenko & Veselov, 2001); and I consider two biased results that can occur during the TT: the Eliza Effect and the Confederate Effect (Warwick & Shah, 2015). And, finally, I discuss the role of Artificial Stupidity in video games, as the privileged scenario in which it is implemented. I review Laird & Duchi (2000), Lidén (2003) and Hingston (2009), who propose a new version of the TT based on video gaming sessions rather than conversational ones, arguing that NPC (non-playable character) entities should be evaluated in terms of “humanness and skill level”, which, I hold, are analogous to ‘human-likeness’ and ‘correctness’.

Along with Jwalapuram (2017), Saygin & Cicekli (2002), and Jacquet et al. (2019), I propose to consider Grice’s Cooperative Principle (1975) as the benchmark to evaluate proper conversations from malicious ones during the TT, such as the conversations exploited by Artificial Stupidity. I consider the four maxims of the Cooperative Principle (Quantity, Quality, Manner and Relation) and I argue that, on the one hand, the maxims of Quantity and Manner probe the style of the replies and are evaluated by ‘human-likeness’; on the other hand, the maxims of Quality and Relation probe the content of the replies and are evaluated by ‘correctness’.

Chapter 3
Blockhead

Abstract. The third chapter concludes Part I of this work. In this chapter, I consider Block's Look-Up Table argument (hereafter LUT), commonly known as Blockhead. Here I review Block (1981); I consider the objections to Blockhead, especially Copeland (2000), Proudfoot (2013) and McDermott (2014); and I propose two other possible versions of Blockhead: (i) Stupid Blockhead, the table of which contains human-like uncooperative interactions (like those involved in Artificial Stupidity); and (ii) Ultimate Blockhead, the table of which is independently learnt rather than hand-coded. Turing (1950) seems to ban Blockhead from the TT, by advocating a learning *child machine* rather than a pre-programmed *adult machine*. I hold, however, that it is logically possible and physically feasible to exploit the experimental design of the TT by brute-force (as Stupid Blockhead shows). One goal of the new version of the TT that I propose is to prevent this from happening.

The Look-Up Table (LUT) argument, against the validity of the TT as a test of intelligence, can be stated as follows:

> **LUT Argument:** *if the responses of an entity are produced by brute-force, then the entity is not intelligent.*

Brute-force is an exhaustive method of problem-solving. In the case of a conversational brute-force approach, it could be described as a systematic enumeration of all the responses with which the system is supposed to reply. Block's (1981) thought experiment Blockhead is an example of a conversational brute-force approach. As described by Copeland (2000), Blockhead:

> “is a hypothetical program able to play the imitation game successfully, for any fixed length of time, by virtue of incorporating a large, but finite, ‘lookup’ table containing all the exchanges with the interrogator that could occur during the length of time in question.” (532)

In the next section, I review “*Behaviourism and Psychologism*” (Block, 1981) and I describe Blockhead. In section 3.2, I show the views against Blockhead. In section 3.3, I discuss a version of Blockhead implementing Artificial Stupidity, which I call Stupid Blockhead. And in section 3.4, I consider an advanced version of Blockhead, able to independently learn its look-up table, which I call Ultimate Blockhead.

3.1 Block’s Blockhead

The first formulation of the LUT argument is given by Shannon & McCarthy (1956), who hold that a definition of ‘thinking machine’ should involve something about the inner workings of the machine:

> “A disadvantage of the Turing definition of thinking is that it is possible, in principle, to design a machine with a complete set of arbitrarily chosen responses to all possible input stimuli [...]. Such a machine, in a sense, for any given input situation (including past history) merely looks up in a ‘dictionary’ the appropriate response. With a suitable dictionary such a machine would surely satisfy Turing’s definition but does not reflect our usual intuitive concept of thinking. This suggests that a more fundamental definition must involve something relating to the manner in which the machine arrives at its responses – something which corresponds to differentiating between a person who solves

> a problem by thinking it out and one who has previously memorized the answer." (v-vi)

Similarly, Block rejects the TT by drawing a comparison between two entities: one of them works by implementing a brute-force approach, the other one works by implementing certain internal information processes. More specifically, Block (1981) argues:

> "The two systems could be exactly alike in their actual and potential behavior, and in their behavioral dispositions and capacities [...], yet there could be a difference in the information processing that mediates their stimuli and responses that determines that one is not at all intelligent while the other is fully intelligent." (5)

Block does not discuss further or specify the inner workings that, according to him, a target entity should be required to possess in order to be attributed with intelligence during the TT. 'Intelligence', it's worth noting, always refers to the kind of intelligence involved in the TT, which is defined by Block (1981) as follows:

> "Intelligence (or more accurately, conversational intelligence) is the disposition to produce a sensible sequence of verbal responses to a sequence of verbal stimuli, whatever they may be". (11)

3.1.1 Block's objections to the TT

The first difficulty raised by Block against the TT is that it is intended to provide an operational definition of intelligence, where 'operational definition' gives an analysis of the meaning of the term 'intelligence'. As Boyd (1983) explains, operationalism was a positivist proposal for rationally reconstructing the use of theoretical terms (that is, those terms which refer to 'unobservables', such as electrons). The point of the proposal was to treat a theoretical term as having an 'operational' meaning, defined in terms of analytic laboratory procedures and measurements that would be used to test its presence and quantify it. As discussed in Ch. 1, I hold that this interpretation is inconsistent with Turing's externalist conception of intelligence. And Block (1981), as well, abandons the operational interpretation in favour of a dispositional one:

> "According to the operationalist interpretation of the Turing Test as a definition of intelligence, it is absurd to ask of a device that passes the Turing Test whether it is *really* intelligent, and it is equally absurd to ask of a device that fails it whether it failed for some extraneous reason but is nonetheless intelligent. This difficulty can be avoided by going from the crude operationalist formulation to a familiar behavioral disposition formulation. On such a formulation, intelligence is identified not with the property of passing the test (if it is given), but rather with a behavioral *disposition* to pass the test (if it is given)." (8)

Nevertheless, according to Block, even if an entity has the disposition to pass the TT, it cannot be considered intelligent unless there is a way to inspect its internal functioning. So, Block designs a logically possible machine with the capacity and the disposition to pass the TT, which uses an unintelligent method nonetheless: at each interaction with its interlocutor, it looks up a gigantic table – or search a gigantic tree – of instructions in order to pick an appropriate reply. This, according to Block (1981), is the conclusive evidence that the entity is lacking intelligence entirely:

> "I conclude that whether behavior is intelligent behavior is in part a matter of *how* it is produced. Even if a system has the actual and potential behavior characteristic of an intelligent being, if its internal processes are like those of the machine described [Blockhead], it is not intelligent." (21) [Italics added.]

The second difficulty that Block raises regards the human judge. According to him, the judge has too much responsibility in the experiment, for she is the one who decides upon the entity's nature and attributes it with intelligence. And since Block (1981) designs an entity which is able to pass the test without intelligence, he is worried by the following question:

> "Could the issue of whether a machine *in fact* thinks or is intelligent depend on how gullible human interrogators tend to be?" (10)

My answer is no. Blockhead, I hold, is not a case of Artificial Stupidity and, therefore, I disagree that it is a matter of the human judge's gullibility. As discussed in the previous chapter, a gullible human judge may be deceived by an 'artificially stupid' entity (either human or machine), which uses certain uncooperative strategies to evade the conversation rather than to hold it properly.[77] Blockhead, on the other hand, is a machine able to look up a gigantic table and to always pick a sensible reply to whatever verbal input, not only to the inputs that are indeed sent, but to the input that might have been sent as well. As Block (1981) clarifies:

> "in order to be intelligent according to the above-described conception, the system must be disposed to respond

[77] Where 'properly' means following the Gricean Cooperative Principle (see Ch. 2).

> sensibly not only to what the interlocutor *actually* says, but to whatever he *might* have said as well.” (ibid)

So, I hold that no human, or any other smarter judge for that matter, would be able to discriminate Blockhead from a human during the TT. Since no human judge could unmask Blockhead, my claim is that it should rather be the experimental design of the TT itself to be improved in order to prevent Blockhead from passing.

Finally, Block (1981) makes this last point against the TT:

> “The fault of the Turing Test […] is one of experimental design, not experimental concept. The trouble is that your Turing Test has a *fixed length*. […] It is certainly true that my machine’s capacity to pass Turing Tests depends on there being some upper bound to the length of the tests.” (34)

I agree with Block when he says that the fault of the TT is one of experimental design rather than one of experimental concept. In other words, the fault of the TT is not in the idea that it is possible to test for intelligence by evaluating a target entity’s verbal competency, but rather in the way along which the entity is parametrised and evaluated. However, I hold that Block points out the wrong fault. He points out that Blockhead’s capacity to pass the TT depends on there being some upper bound to the length of the tests, in the same way as the human capacity to pass the TT does. In other words, Blockhead would fail an unlimited TT by eventually running out of space in its table (which is huge but finite), just as a human would fail by eventually passing away before the test ends. The unlimited TT, because of such impracticability, does not seem a proper update for the TT in order to rule out Blockhead from passing. Since the unlimited TT would be failed not only by Blockhead but by any human being as well, I argue that a different attempt to update the test is required. A preferable updated version of the TT would thus be a test which any human could pass, but that Blockhead could not.

Moreover, despite Block's argument, even a time-limited TT seems dangerous for Blockhead, since it cannot always be able to work fast: given the size of its table, even if it were able to search at light-speed, it could potentially take years to emit a reply. As Copeland (2000) remarks:

> "firstly, the proposed recipe for building a brain-emulator cannot work, given practical limitations on storage capacity; and secondly, even if this point is set aside and we suppose that such a machine [Blockhead] were actually to be constructed, it would *not* emulate the brain, since what the brain can do in minutes would take this machine thousands of millions of years." (533)

3.2 Objections to Blockhead

There are a number of objections to Blockhead. The two most common are: (i) that Blockhead is only logically possible, and not physically feasible, and therefore it does not undermine the TT; and (ii) that an entity like Blockhead would indeed possess mental states and should be considered intelligent.

3.2.1 Proudfoot and Copeland

Proudfoot (2013) rejects Blockhead for two reasons. The first, is that it is incompatible with Turing's externalist conception of intelligence. According to the response-dependent reading of Turing, it is not the entity's behavioural capacities – or even dispositions – that determine whether it is intelligent. It is, to use the words that perhaps Turing would have used, the "interrogator's emotions."[78] On the one hand, Turing does not discuss the proper states of mind that a judge should have (suffice to say that the judge should be as much as possible unbiased). On the other hand, he defines the "*properties of the object under consideration*"[79] by describing the kind of machine that should be built and tested for intelligence: a "child machine"[80], that is, an *unorganised machine* which is able to learn. In section 3.2.3, I discuss the difference between *child* and *adult machines* in more detail.

The second reason why Proudfoot (2013) rejects Blockhead is due to its physical unfeasibility:

> "Even if Turing's criterion is a logically sufficient condition, it is not undermined by Block's example. Based on the same remarks in Turing's 1952 broadcast, we can, I suggest, read his criterion as follows: a machine is intelligent if actually it passes the Turing test. Using the "Actually" operator A, AP is true iff P is true in the actual world; […] Reading Turing's criterion as a logically sufficient condition that incorporates the "actually" operator fits not only with his emphasis on real-world machines but also with the response-dependence interpretation of his test." (401)

Similarly, Copeland (2000) rejects the LUT for being a logical possibility only:

[78] See Turing (1948).
[79] Turing (1948, p. 19).
[80] Turing (1950, p. 456).

> "If Turing had been proposing a definition of 'thinking' – a logically necessary and sufficient condition – or even merely a logically sufficient condition, then the lookup table objection would indeed count against the proposal […]. However, there is no reason to believe that Turing was claiming anything more than that his principle[81] is actually true. The other-worldly possibility of a lookup-table machine that is fast enough to emulate the brain has no tendency at all to show that Turing's principle is *actually* false. Likewise, it is no challenge to the actual truth of the Church-Turing thesis that a human rote-worker who occupies a possible world in which the human memory is unlimited can be in the process of writing down a number that is not computable by any Turing machine (see Turing, 1936: 231, 249–252)." (533)

3.2.2 McDermott

The second line of arguments is advocated by McDermott, who attempts to reject Block's thought experiment (which he calls HTA, Humongous-Table-Argument) not by focusing on the physical unfeasibility of Blockhead (which he calls HTP, Humongous-Table-Program), but instead by arguing that Blockhead in fact possesses intelligent features. McDermott (2014) argues that, first, for an HTA to work:

> "two aspects of the test are crucial:
>
> - There must be a fixed time limit for the conversation between judge and interlocutor. Otherwise a simple table wouldn't be adequate.

[81] "Turing's Principle: A machine that by means of calculation imitates – or, better, 'emulates,' for Turing is concerned with faithful imitation – the intellectual behaviour of a human brain can itself appropriately be described as a brain, or as thinking." (Copeland, 2000 p. 529).

- The judges must not be able to compare notes before deciding whether an interlocutor is human. Otherwise the fact that the same table governs all the conversations might give the game away." (148)

The first premise is true for humans as well: "fixed time limit" means that the conversation in the TT cannot be infinite but should rather have a reasonable duration for the entity – either human or machine – to hold it. The second premise seems to imply that the look-up table would give the same response over and over again to the same verbal stimulus. If this is true then there could be easier ways to unmask Blockhead than to compare previous conversations: for instance, one could simply ask the same question twice, or ask the entity to recall a piece of information given previously during the conversation. It is not so clear, however, if it would be the case: Blockhead's tree has, for each interaction, sub-branches with new replies (such as "you already asked me that"). So, if a question is repeated, the HTP can pick uncompromising responses, it only needs to keep track of the questions asked.

In order to legitimately consider an HTP intelligent, McDermott (2014) recognises three prerequisites:

- "That the HTP must be exhaustive, and not be based on some vaguely imagined set of tricks.
- That the HTP must not be created by some set of sentient beings enacting all possible responses.
- That in the current state of cognitive science it must be an open possibility that a computational model of the human mind will be developed that accounts for at least its nonphenomenological properties." (143)

The first prerequisite suggests that, in order to be considered intelligent, the HTP must not resort to Artificial Stupidity (see Ch. 2); I discuss further this position in section 3.3, where I propose another possible version of Blockhead, which I call Stupid Blockhead. The second prerequisite suggests that, in order to be considered intelligent, the HTP must not be hand-coded by humans; I discuss further this position in section 3.4, where I propose a third possible version of Blockhead, which I call Ultimate Blockhead. The third prerequisite, finally, is meant, as McDermott (2014) writes, to

> "take all issues concerned with phenomenal consciousness off the table [… in order to] allow for the possibility of finding a computational model of human thought that accounts for everything else." (164)

I agree with McDermott that entities using the uncooperative strategies of Artificial Stupidity, as well as entities with a built-in knowledge based, should be banned from the TT. However, I disagree that a cooperative Blockhead which independently acquires its look-up table (like Ultimate blockhead, see section 3.4) ought to be attributed with intelligence. The reason is that the method behind its functioning would still be a brute-force one.

3.2.3 Child machines Vs adult machines

It can be argued that Turing implicitly rules out Blockhead by ruling out *adult machines* in favour of *child machines,* where the former belongs to the set of brute-force LUTs and the latter to the set of learning systems. Turing (1950) suggests the following:

> "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's?" (456)

So, whereas the *adult machine* has a hand-coded, built-in knowledge base, the *child machine* has no pre-programmed information: it is an *unorganised machine* with the ability to learn, similarly to a human child. And like a human child, who requires teaching to become adult, the *child machine* needs an education to "grow" and know things. The idea of the *child machine* is initially developed in Turing (1948), where he summarises the machine he has in mind as follows:

> "The possible way in which machinery might be made to show intelligent behaviour are discussed: […] it is pointed out that the potentialities of the human intelligence can only be realised if suitable education is provided. The investigation centres around an analogous teaching process applied to machines." (431)

There are two possible readings of "*the potentialities of the human intelligence can only be realised if suitable education is provided*." One is that human intelligence is realised through suitable education; the other is that human-level intelligence (even in non-human entities) can only be realised through the same process of education. The first reading seems safe enough: humans would not be able to develop complex cognitive faculties (such as speaking and understanding a language) by their own, with no cultural or social interactions with other human fellows. The second reading is more radical: it implies that intelligence cannot be produced, it can only be acquired. Turing introduces learning machines as machines that are able to modify their own instructions independently, rather than by means of direct human interference. As Turing (1948) explains, it is possible to

> "distinguish two kinds of interference. There is the extreme form in which parts of the machine are removed and

> replaced by others. This may be described as "screwdriver interference". At the other end of the scale is "paper interference" which consists in the mere communication of information to the machine, which alters its behaviour. […] When it is possible to alter the behaviour of a machine very radically we may speak of the machine as being 'modifiable'. […] One may also sometimes speak of a machine modifying itself, or a machine changing its own instructions. […] We shall mainly be interested in paper interference […]." (8)

Now it is possible to interpret Turing's view on how an intelligent machine should be built in two ways. First, Turing is not committing to any particular strategy for building AI: he simply favours the *unorganised child machine* project over the *adult machine* one, without excluding an hand-coded *adult machine* (like Blockhead) from being intelligent. Second, Turing is saying that intelligence could only be ascribed to learning systems, suggesting thus the ability to learn as a necessary condition for intelligence.

In support of the first interpretation, Turing (1950) explicitly points out that there may be disagreement on how an intelligent machine should be built and on what it should be able to do; and, so, that a number of approaches should be attempted:

> "Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and

> named, etc. Again I do not know what the right answer is, but I think both approaches should be tried." (460)

It is worth noting that, in the passage above, Turing seems to suggest that one way to learn a language is to build an embodied entity that can experience the world, foreseeing the subcognitive argument provided by French (1990, 2000a, 2000b) and even anticipating the Total Turing Test proposed by Harnad (1989, 1991, 2000).

On the other hand, Proudfoot (2017) supports the second interpretation:

> "For Turing, learning is the key to intelligence […] the 'hallmark of intelligence is the ability to learn' and, like 'a newborn baby', a computer's possibilities depend upon the education which is fed into it." (317)

Proudfoot (2017) also criticises Expert Systems, programs that are able to show the intelligent behaviour typical of a human expert in a certain endeavour; and she points out that such an approach conflicts with Turing's proposal to build an *unorganised child machine*. She distinguishes between two possible kinds of intelligent machines:

> "We must build a machine that is a "person" with sufficient language understanding to be educable, both by example and by precept. The goal of AI should be, not only a human-level machine, but a human-type machine." (318)

According to this view, Blockhead is a human-level machine (and, likely, even a superhuman-level machine), in terms of how it behaves. However, it is not a human-type machine, in terms of how it learns to behave. In other words, Blockhead is able to hold an open conversation at the same level as a human, but the type of its internal processes is different from that of humans. It is necessary, however, to highlight that Turing, faithful to the idea of Multiple Realizability, never ruled out either machines with internal processes similar to the human ones or machines with internal processes different from the human ones. And the main reason is that we do not know – at least, yet – what the internal processes from which human intelligence emerges are. So, the machine's imitation does not involve the reproduction of human inner workings, but rather the reproduction of human behaviours. As Sprevak (2017) points out, Turing's idea:

> "is an exercise in weak modelling. His aim is to show that Turing machines and human clerks solve the same class of problems: they are capable of producing the same pattern of behaviour. His argument requires him to show that a Turing machine can copy the behaviour of the clerk and vice versa (weak modelling). It does not require him to show that the Turing machine reproduces that clerk's internal psychological mechanisms for generating his behaviour (strong modelling)." (283)

3.2.4 Symbols in, symbols out

My reply to Blockhead is focused on the context in which Blockhead works well. Blockhead is designed to hold an open one-way conversation: a back-and-forth exchange, that is, when the interlocutor sends an input first and Blockhead replies. This system can be defined, as Harnad (1989, 1990) proposes, with the acronym SISO, which means 'Symbols In, Symbols Out'. Blockhead needs the interlocutor's interactions first to produce some *appropriate* output (otherwise it would just emit entries picked *randomly* from its table). In other words, Blockhead needs to receive a verbal input first in order to search its database, find the matches and pick the appropriate response. As Ben-Yami (2005) puts:

> “Because of the way it was programmed, the machine emits only ‘sensible’ strings, strings in which its contribution is always sensible. It would, therefore, pass any test that is *limited to conversational output* [not input] and which determines that a conversant is intelligent only according to behavioral criteria—linguistic criteria, in this case.” (182) [Italics added]

Therefore, I hold the following generalisation: a brute-force LUT, parametrised along ‘human-likeness’ and even ‘correctness’, can only work in the SISO setup, that is, when the test is limited to conversational outputs only. To prevent brute-force strategies, the solution I propose is to modify this setup. In order to modify the SISO setup in the case of the TT, I claim that it is necessary to modify the task of the test, that is, to hold an open conversation. My proposal is to change the task of the test from holding an open conversation to accomplishing an enquiry, and to switch from SISO to SOSI (Symbols Out, Symbols In), In the SOSI test I propose, the entity needs to send a verbal stimulus first (a question) in order to receive a verbal response, needed to accomplish the aim of the enquiry. The SOSI setup allows us to parametrise the hidden entity along a further dimension in addition to ‘human-likeness’ and ‘correctness’: what I call ‘strategicness’, which is intended to rule out brute-force strategies and prevent thus Blockhead from passing the test. I will discuss the dimension of ‘strategicness’ in more detail in Ch. 5, where I present the QTT, the new SOSI test of intelligence that I designed on the model of the TT, with the purpose to avoid the two main difficulties with the experimental design of the TT discussed so far: Artificial Stupidity and Blockhead.

Summing up, the core of the LUT argument lies in the claim that intelligence does not supervene on behaviour (be it behavioural capacity or disposition). Blockhead possesses a look-up table, hand-coded by humans, which enables it to pair every possible verbal input to an appropriate verbal output. The internal processes of Blockhead are thus very different from the human ones, for they do not depend on previous thoughts, inclinations, beliefs, desires, and so on. I agree that Blockhead represents a difficulty for the experimental design of the TT, rather than its experimental concept[82]. But I do not think the solution is to train the judge, or argue that Blockhead should be considered intelligent[83]; I rather argue that the experimental design of the TT should be improved in order to avoid potential exploits, like Artificial Stupidity, and brute-force approaches, like Blockhead. My goal is to propose a new test immune to Blockhead, as well as to a human with no verbal intelligence, who could pass the TT by memorising, thanks to her extraordinary memory, any possible verbal interaction whatsoever.

3.3 Stupid Blockhead

As I pointed out in the previous section, Blockhead could pass the TT by means of an unintelligent method, undermining thus the reliability of the TT's results. However, Blockhead is only logically possible, meaning that it is not possible to build it; and even if it were built, it would be extremely slow (jeopardizing its success). To counter this objection, I argue that Blockhead could be more easily equipped with the uncooperative strategies of Artificial Stupidity (see Ch. 2) in order to exploit the judge's beliefs and pass the test, with no need of a complete table to look up. As Block (1981) remarks:

[82] See Block (1981, p. 34).
[83] See McDermott (2014).

> "If one sets one's sights on making a machine that does only as well in the Turing Test as *most* people would do, one might try a hybrid machine, containing a relatively small number of trees plus a bag of tricks of the sort used in Weizenbaum's program." (34)

So, Block (1981) distinguishes between two types of LUT. One is – the very slow – Blockhead, which is able to hold a human-like open conversation by looking up its complete table of all the possible verbal interactions, and by always picking an appropriate response for whatever verbal stimulus. The other is the Elizish Blockhead, or as I call it, Stupid Blockhead, which is able to hold a human-like open conversation by looking up its incomplete table of interactions, and to its table of uncooperative tricks whenever is needed. So, Stupid Blockhead is an incomplete version of Blockhead, with an incomplete table of appropriate responses. It can play the TT as the original Blockhead until it receives an input with no correspondence in its table, whereupon it can play the TT as the artificially stupid Eliza, using certain uncooperative strategies to bring back the conversation to some manageable topic. Or it may as well be a pure Stupid Blockhead, with no table of appropriate interactions at all, which keeps evading the conversation all the time by being uncooperative, as long as in a human-like way (an extreme version of Stupid Blockhead could only contain one entry: "I plead the fifth"). Stupid Blockhead, given its limited table, is not only logically possible, but also physically feasible.

It is worth noting that it would be plausible for a human to use evading strategies to avoid unwelcome conversations. So why should Stupid Blockhead or Artificial Stupidity be regarded as an exploit? My answer is that humans learn such tricks and strategies, they don't memorise them from a book; and those tricks and strategies are certainly not pre-coded in our DNA. Analogously, if the entity is – to use the word that

Turing (1950) chooses – “educated”, that is, if either the human baby or the *unorganised machine* learn to produce certain behaviours (including the criticised artificially stupid behaviours), then those behaviours would be regarded as intelligent, no matter their uncooperativeness.

3.4 Ultimate Blockhead

Following Turing’s (1950) proposal of an *unorganised child machine*, here I propose to consider a machine which is able to learn to become Blockhead, which I call Ultimate Blockhead.

Recalling what is intended with *child machine*, Turing (1950) proposes the analogy between a blank machine and a new-born child, both of which can initially do little by their own but, after undergoing an education process, can learn to do what an adult human can do. As Sterrett (2012) remarks:

> “This analogy […] provides Turing with the means to respond to one of the most common objections raised against the possibility that a machine could be regarded as exhibiting intelligent behavior. This objection (to the possibility of intelligent machinery) is, in Turing’s words, the view that “[i]nsofar as a machine can show intelligence this is to be regarded as nothing but a reflection of the intelligence of its creator.” (704)

Ultimate Blockhead is a machine which is able to do anything Blockhead and Stupid Blockhead can do, with the difference that the latter ones use pre-programmed, hand-coded tables, whereas the table of Ultimate Blockhead is independently acquired, or learnt. Ultimate

Blockhead learns its table by, let's say, scouring the whole internet (let's suppose it has access to every bit of information contained on the web) and memorising every verbal interaction whatsoever. It can be said that whereas "the whole point of the machine [Blockhead] is to substitute memory for intelligence,"[84] the whole point of Ultimate Blockhead is to substitute memorisation for learning.

So, should Ultimate Blockhead be considered intelligent? Ben-Yami (2005) would say yes. Whereas Block (1981) stresses out that the Blockhead's lack of intelligence is due to its internal organisation as a LUT and its inner workings, Ben-Yami (2005) identifies the problem in its learning history. According to this view, it is how Blockhead's table is acquired that matters: if it is hand-coded by humans, then it lacks intelligence; if it is independently acquired, then it is intelligent. So, Ben-Yami (2005) agrees that Blockhead lacks intelligence, but

> "contrary to Block, this is not determined by reference to its internal processes. The machine lacks intelligence because all it does is reproduce answers that were given to it in advance. And this is determined by reference to the relation between the answers that were formerly given to it and to the answers it now gives. The machine is not intelligent for the same reason that Christian is not a poet: Christian answers Roxane what Cyrano tells him to answer, and the machine answers its interrogator what the programmers 'told' it to answer." (183-184)

This would not be the case for Ultimate Blockhead, which would be able to learn its table independently rather than receiving it from the programmers.

[84] Block (1981, p. 34).

However, unlike humans, Ultimate Blockhead can learn impressive amounts of information in a short time. Let's say that from the moment in which Ultimate Blockhead is turned on, it takes one year to scour the whole internet and memorise every interaction whatsoever. So Ultimate Blockhead doesn't really need an educator, as Turing (1950) suggests for the *child machine*, since the kind of learning that Ultimate Blockhead requires, that is, memorisation, can be automated. Nonetheless, Ultimate Blockhead can still be viewed as a *child machine*: initially it is an *unorganised machine* similar to an infant cortex – but with an exceptional memory –, unable to hold even the simplest of conversations; and after some time spent memorising the verbal interactions found on the internet, it becomes very good at it. Once its table is acquired, Ultimate Blockhead can do, on average, as good as Blockhead in the TT. I say *on average* since the human judge has certain ways to corner Ultimate Blockhead, for instance by using neologisms never ever used on the web. It's worth noting that the original Blockhead would, on the contrary, *always* be able to produce an appropriate output, neologisms or entirely new languages included.

It is also worth noting that the learning ability of Ultimate Blockhead is limited: for instance, it cannot learn during the TT, that is, it is not able to learn something the judge teaches it (e.g. a neologism that never appears on the web). Most importantly, Ultimate Blockhead is not able to learn new information from induction or deduction. The reason is that Ultimate Blockhead does not possess a general learning ability, like humans, but just the learning ability of an aggregator: all it does is memorising.

So, to rephrase the question asked above, "should Ultimate Blockhead be considered intelligent?", should Ultimate Blockhead pass the TT? My answer is no: even though the table of Ultimate Blockhead is learnt and not hand-coded, it still works like the original Blockhead, that is by brute-force. This is not compatible with the generalisation of the LUT argument that I proposed at the beginning of the chapter: *if the responses of an entity are produced by brute-force, then the entity is not intelligent*. And, as Turing (1950) speculates:

> "Intelligent behaviour presumably consists in a departure from the *completely disciplined* behaviour involved in computation." (459) [Italics added]

If this is true, then neither Blockhead nor Ultimate Blockhead (and certainly not Stupid Blockhead) should be considered intelligent and, consequently, they should not be allowed to pass the TT. As Ben-Yami (2005) specifies:

> "Other variants of Block's machine can also be imagined, in which, although the internal information processing is again different, the relation between linguistic input and output is the same or similar, and consequently these machines are also devoid of intelligence." (184)

Summing up, in the case of the original Blockhead proposed by Block (1981), the table is hand-coded by humans; in the case of Ultimate Blockhead, a new LUT that I propose, the table is acquired independently, that is, learnt or, better, memorised. What should be kept in mind, however, is that its learning ability is limited to the ability of an aggregator, to look for new information and to compile them. In other words, for Ultimate Blockhead, learning means memorising interactions – found, let's say, on the internet – and reusing them appropriately. It does not possess the general learning ability of a human child, who can learn many different things in many ways. Moreover, both Blockhead and Ultimate Blockhead use brute-force.

Consequently, as for any LUT, they should not be considered intelligent and, thus, be prevented from passing the TT.

To do so, I propose to evaluate *how well* the target entity performs during the TT, by evaluating a further dimension in addition to ‘human-likeness’ and ‘correctness’: what I call ‘strategicness’. I will discuss it in more detail in Ch. 5 and 7, where I propose the QTT, a new test of intelligence on the model of the TT which is able to avoid not only Blockhead and Stupid Blockhead (as well as Artificial Stupidity in general), but also Ultimate Blockhead.

3.5 Summary

In this chapter, I discuss Blockhead (Block, 1981), a brute-force system which is able to pass the TT by looking up its table and picking an appropriate response to whatever verbal stimulus it receives. I hold that Blockhead, no matter the arguments appealing its physical unfeasibility (Copeland, 2000; Proudfoot, 2013) or its learning history (Ben-Yami, 2005; McDermott, 2014), undermines the reliability of the TT’s results.

I propose two other versions of Blockhead, one physically feasible, and one with a learning history: (i) Stupid Blockhead, which can pass the test by using the Artificial Stupidity’s strategies to exploit the judge’s beliefs; and (ii) Ultimate Blockhead, which independently acquires its table by memorising all the possible verbal interactions on a given source (e.g., the internet). Since Blockhead, Stupid Blockhead and Ultimate Blockhead work in the same way, that is, by brute-force, I argue that none of them should be attributed with intelligence; and that the experimental design of the TT should be improved in order to rule out LUT entities from passing.

In the second part of the thesis, I introduce my proposal for a new conversational test of intelligence, which I call the Questioning Turing Test, with the goal to avoid both Artificial Stupidity and Blockhead (as well as Stupid and Ultimate Blockhead). In order to prevent brute-force entities from passing the TT, I propose a further dimension along which to evaluate the entity in addition to 'human-likeness' and ' correctness': what I call 'strategicness'. To evaluate 'strategicness', I switch from the SISO setup of the TT (where the task is to hold an open conversation) to the SOSI setup of the QTT (where the task is to accomplish an enquiry with as few human-like questions as possible).

Part II

Chapter 4

Alternative and Extended Versions of the Turing Test

Abstract. The second part of the thesis is dedicated to my proposal of a new version of the TT, that I call the Questioning Turing Test (QTT), intended to prevent Artificial Stupidity (Ch. 2) and Blockhead (Ch. 3) from passing. In this chapter, I discuss two alternative versions of the TT: the Minimum Intelligence Signal Test (MIST), proposed by McKinstry (1997, 2009); and the Feigenbaum Test (FT), proposed by Feigenbaum (2003). The MIST addresses the problem of the subcognitive competency, discussed by French (1990). The FT addresses the problem of the judge's expertise, is intended to evaluate Expert Systems. Both the MIST and the FT provide some ground for the experimental design improvements that I propose with the QTT. I also discuss two extended versions of the TT: the Total Turing Test (3T), proposed by Harnad (1989, 1991, 2000); and the Truly Total Turing Test (4T), proposed by Schweizer (1998, 2012a).

So far, I described the Turing Test (Ch. 1), and I discussed what I hold to be the two main difficulties with it: Artificial Stupidity (Ch. 2) and Blockhead (Ch. 3).

In this chapter, I discuss two alternative versions and two extended versions of the TT. The alternative versions are intended to address a specific problem with the experimental design of the TT. The two alternative versions of the TT that I discuss are (i) the MIST, proposed by McKinstry (1997, 2009) and intended to probe the subcognitive competencies of an entity; and (ii) the FT, proposed by Feigenbaum (2003) and intended to evaluate an entity's ability to display and justify expert knowledge. The MIST limits the interactions to yes/no ones, and rules out Artificial Stupidity by parametrising the entity along 'correctness' rather than 'human-likeness'. The FT evaluates (i) the entity's use of the language ('human-likeness'), (ii) the entity's rightness and cooperativeness ('correctness'), and (iii) the entity's ability to provide justifications and explanations for its replies ('strategicness'). The extended versions are intended to broaden the experimental design of the TT, limited to verbal behaviours. The two extended versions of the TT that I discuss are (i) the 3T (Harnad, 1989), which includes all the possible sensorimotor behaviours of a human, not only verbal ones; and (ii) the 4T (Schweizer, 1998), which includes the evolutionary mechanisms to replicate the human achievements.

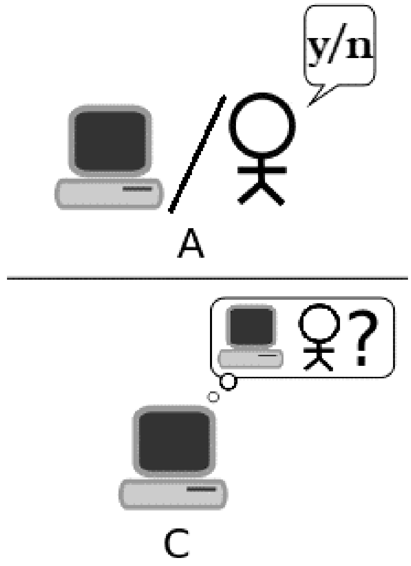
In the next section, I present McKinstry's quantitative and automatable test of intelligence: the MIST. In section 4.2, I describe Feigenbaum's Expert Systems and the FT, an expert subject matter TT. In section 4.3, I examine the extended versions of the TT: the 3T, proposed by Harnad (1989, 1991, 2000); and the 4T, proposed by Schweizer (1998, 2012a).

4.1 The Minimum Intelligence Signal Test

Inspired by French's (1990) argument on the subcognitive competencies (discussed in Ch. 1), McKinstry (1997, 2009) designs the Minimum Intelligence Signal Test (MIST), intended as an objective and automatable version of the TT. The primary purpose of the MIST, as McKinstry (2009) states, is to provide a possible:

> "method for the automatic discovery of a universal human semantic-affective hyperspatial approximation of the human subcognitive substrate – the associative network which French (1990) asserts is the ultimate foundation of the human ability to pass the Turing Test – that does not require a machine to have direct human experience or a physical human body." (283)

The MIST is conducted in the following way: a judge (C) – which can be a machine – asks yes/no subcognitive questions to the target entity (A) in order to decide whether the entity is human or machine (as shown in fig. 1). The judge's decision depends entirely on how many correct binary answers the entity gives, against the probability that the entity's replies are given randomly.



(fig. 1)

As McKinstry (1997) specifies:

> "Given a series of stimuli (items), a system being tested generates a binary response for each stimulus. Thus a Minimum Intelligent Signal may be detected in the cumulative binary output of that system. A system that has a MIST score that is statistically different from a random system is said to be intelligent. A system that has a MIST score that does not differ statistically from the MIST score of an average human, is said to have the intelligence of an average human." (17)

The MIST follows Turing and French in the claim that a conversational test is adequate to test for intelligence, but its experimental design introduces a few restrictions. The entity in the MIST is not required to be able to hold a conversation: it can only reply yes or no to the binary subcognitive questions asked by the judge. For this reason, McKinstry (1997) argues, it is not even necessary that a human judge carries the questioning: the questions can be automatically delivered to the system. That is why, as McKinstry points out, the MIST is intended to be an automatable version of the TT, and to provide statistical evidence of the machine's general knowledge and common-sense (by means of French's subcognitive questions). The corpus of subcognitive items (propositions) and their Boolean answers is called Mindpixels: the idea behind the name is that, similarly to a picture made of pixels, the more the pixels, the higher the quality of the picture, a – conversational – mind is made of propositions, the more the propositions available, the more human-like the mind. McKinstry launched the Mindpixels project in 2000, with the aim to create a general and common-sense knowledge base made of millions of human yes/no statements. In four years, by 2004, the Mindpixels server registered 1.4 million entries. As McKinstry (2009) points out, even though the Mindpixels corpus cannot be considered intelligent itself, it

> "would allow for the high speed automation of the Turing Test and aid in the discovery of a truly intelligent computer program." (290)

Given its experimental design, the MIST is able to rule out Artificial Stupidity, since the interactions during the conversation are limited to yes/no ones. For this reason, the MIST does not parametrise the entity along 'human-likeness', but only 'correctness'. In other words, the MIST factors out the style of the replies, preventing the target entity from exploiting the judge's beliefs. McKinstry (1997, 2009) distinguishes four distinct stages, that can be summarised as follows:

(i) Collect and validate a corpus of binary propositions which require human intelligence to respond to and which have a stable response. (For instance, "my foot is bigger than the Sun" has a stable human response: false; "my foot is bigger than yours" does not have a stable response).

(ii) Draw a test of at least 20 random propositions from the corpus, such that 50% are true, and 50% are false.

(iii) Present these propositions to the target entity in random order.

(iv) Calculate the probability that the system is human or random (machine).

In the case of an entity scoring 15/20 of correct propositions every time it is presented with 20 items randomly selected from a very large pool, it is possible to say that the entity is statistically human. In the case of 15 correct propositions out of 20, the probability that the entity is a random system (that is, a coin) is only .04; in the case of an entity scoring 20 correct propositions out of 20, the probability that the entity is a random system is now less than .0002, specifically 1 in 2^{20}, that is, 1 in 1.048.576 (the same probability to flip a coin either heads or tails 20 times in a row). It can be argued that he experimental design of the MIST does the opposite than the experimental design of the TT: whereas in the latter 'human-likeness' is conflated with 'correctness', in the former 'correctness' is conflated with 'human-likeness'. In other words, the MIST infers 'human-likeness' from 'correctness'.

Notably, McKinstry (1997) distinguishes between cooperative and evasive replies when referring to, respectively, consistent or inconsistent replies with human intelligence. As McKinstry (1997) writes:

> "For each item, judge the item/response pair either consistent or inconsistent with human intelligence. This grading procedure may be easily automated, reducing the chance of grading error or unforeseen bias. Sum the total of the items judged consistent I (Intelligent), and sum the items judged inconsistent E (Evasive). Probability the system under consideration is intelligent and cooperative is p(I)=I/N. Probability the system is intelligent and evasive is p(E)=E/N. Both probabilities must sum to 1.0." (18)

This distinction, I argue, is compatible with the distinction I made in Ch. 2 between the cooperativeness of an intelligent conversational entity and the evasiveness of a 'stupid' conversational entity. For this reason, and since the interactions are limited to yes/no questions (and, consequently, the style of the replies is factored out), I argue that the experimental design of the MIST prevents Artificial Stupidity from passing. Given the design of the MIST, an artificially 'stupid' entity could not use the language to exploit the judge's beliefs; and it would have no more than a .0002% chance of answering correctly by chance, being thus considered intelligent. So, the case in which an entity – either human or machine – can win exclusively by chance is ruled out, given the very low probability of scoring 20 out of 20 correct replies by guesswork. And, it's worth noting, even if an unintelligent system wins the MIST with a score of 20/20 by chance, it still would not be enough to claim that the system is intelligent, since it would be required to match its performance in future tests (which is statistically very unlikely). An entity guessing the answers randomly, being the possible answers “yes” or “no” only, would averagely score .5 in the test (replying correctly to 10/20 of the subcognitive questions). So, given the nature of the questions, and the low probability of answering all of them correctly, if the entity ends up with a MIST score of >.5 over a series of sessions, it is statistically possible to infer that the entity shows intelligence. It is important to underline that in order to test an entity properly, both French and McKinstry seem to suggest that the judge should be a trained one rather than an average person. In McKinstry’s (2009) words, “the Mindpixel system is a many-headed version of French’s competent and prepared investigator”[85], that is, a judge who asks subcognitive questions in order to unmask the target entity. The MIST, finally, can be made more and more difficult depending on how many questions the entity must answer. As McKinstry (2009) puts it:

> “the chance of coin flipping appearing as a perfectly intelligent person in an arbitrary 1,000 items MIST would be one in 2^{1000} – a very, very small chance indeed.” (289)

Summing up, the Mindpixels online corpus and McKinstry’s MIST are intended to deal with French’s notion of subcognitive competency, which is argued to allow the judge to open a window on both the unconscious cognitive structure and the physical characteristics of an entity. The MIST is a quantitative test for human-like intelligence, statistically measuring the amount of correct yes/no answers to 20 subcognitive questions: the idea is that a system able to reply correctly to more than 10 out of 20 subcognitive yes/no questions, over a series of sessions, would provide a statistical proof of its intelligence. The judge in the MIST, as the judge in French’s Rating Games, is required to be a trained one (in asking subcognitive questions). In the MIST, moreover, the role of the judge does not involve any direct decision about the hidden entity’s ‘human-likeness’. The target entity is

[85] McKinstry (2009, p. 291).

parametrised along one dimension: 'correctness'. It is not parametrised along the dimension of 'human-likeness', since the interactions allowed are binary, unstylish ones only. Because of the limited interactions available, the test could be automated, with the human judge simply replaced by a machine. The MIST is intended to evaluate the patterns of an entity's replies, making it possible to gain statistical evidence. It is possible to say that whereas in the TT 'human-likeness' is conflated with 'correctness', in the MIST 'correctness' is conflated with 'human-likeness'. Here, intelligence is inferred from 'human-likeness' which is, in turn, inferred from 'correctness': the more the correct replies given by the entity to the judge's binary subcognitive questions, the more human-like its performance. Finally, the experimental design of the MIST, by parametrising the hidden entity along the dimension of 'correctness', can prevent Artificial Stupidity from passing, but not Blockhead[86].

4.2 The Feigenbaum Test

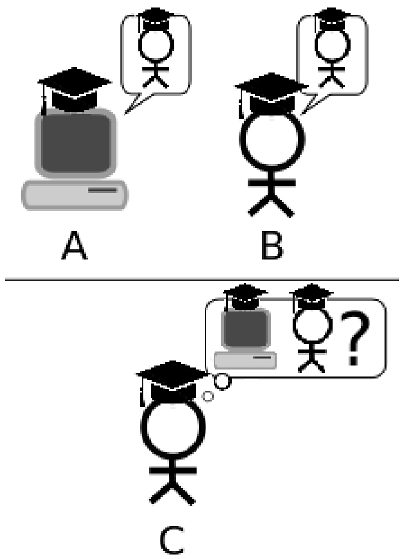
Focused on Expert Systems, Feigenbaum (2003) proposed a version of the TT where (i) the task is to hold an expert dissertation rather than an average conversation; and (ii) the judge is an academic expert rather than an average person.[87] The Feigenbaum Test (FT) is designed to evaluate different abilities, as Feigenbaum (2003) points out:

> "(i) The ability to concatenate assertions and arrive at a new conclusion. [...]
> (ii) The ability to learn and use external symbols to represent numerical, spatial, or conceptual information.

[86] As a matter of fact, the Mindpixel corpus is a Blockhead.
[87] Where "average" simply means without any specific competency or any particular deficiency, in order to rule out false positives and negatives (see Proudfoot, 2013).

(iii) The ability to learn and use symbols whose meanings are defined in terms of other learned symbols.
(iv) The ability to invent and learn terms for abstractions as well as for concrete entities.
(v) The ability to invent and learn terms for relations as well as things." (33)

The FT is played by three participants (as shown in fig. 2): an expert entity, let's say in molecular biology (A); a human expert in the same domain, let's say a molecular biology professor member of the National Academy (B); and a human expert judge, let's say another molecular biology professor member of the National Academy (C).



(fig. 2)

The purpose of A is (i) to hold an expert conversation with C as good as – or better then – B and (ii) to be recognised as human as often as B. The language allowed is restricted to the academic jargon of a certain discipline, so rather than general 'human-likeness', the FT evaluates what I call 'expert human-likeness'. As Feigenbaum (2003) clarifies:

> “The judge poses problems, asks questions, asks for explanations, theories, and so on—as one might do with a colleague. Can the human judge choose, at better than chance level, which is his National Academy colleague and which is the computer?” (36)

The dimensions along which the entity is parametrised in the FT are (i) ‘expert human-likeness’ and (ii) ‘correctness’. That is, the abilities of the target entity to, respectively, be recognised as a human and produce valid dissertations, discoveries, justifications (and so on). I argue, however, that there is another, implicit dimension along which the entity is evaluated: ‘strategicness’, that is, the method with which an entity argues, discovers, justifies its conclusions (and so on). Because of those dimensions, I argue that the FT has three advantages.

The first is that the FT rules out superhuman intelligence, by parametrising the entity along ‘expert human-likeness’. Feigenbaum (2003) considers the case of an infallible Expert System, which he calls the Ultra-Intelligent Computer (UIC). Similarly to the *oracle machine* proposed by Turing (1939), Feigenbaum (2003) notes that the UIC would fail the FT, because of its non-human-like infallibility, from which it would be possible to infer its ‘human-unlikeness’:

> “Paradoxically, the UIC would be easily discernible from the elite human performer. It would be offering inductions, problem solutions and theories that were not yet reached by any human, yet were plausible, rigorous upon explanation, and either correct or interesting enough to subject to experimental test.” (36)

The second advantage is that the FT rules out Artificial Stupidity, since it evaluates the entity specifically for its ability to produce pertinent and correct responses. Moreover, the jargonized language allowed in the FT prevents Artificial Stupidity from passing the test by exploiting the judge's beliefs, unlike in the TT, where the language is not restricted, and evasive tricks are possible. As Feigenbaum (2003) specifies:

> "To factor out the facet of intelligence related to full natural language [use and] understanding, the framers of the game might decide at the outset that the interactions will be conducted in the heavily jargonized and stylized language with which practitioners in the selected domains usually communicate their problems and issues." (*ibid*.)

The third advantage is that the FT can rule out Blockhead as well, which would not be able, given its inner workings, to "concatenate assertions and arrive at a new conclusion"[88], or to "invent and learn terms for abstractions."[89]

There are two problems with the FT. First, it is an impractical test: it not only requires to recruit expert humans, but also to build such an advance expert machine. Second, 'strategicness' is not explicitly defined, and the experimental design of the FT does not allow to measure it. 'Strategicness' still plays an important, although implicit, role in the FT: for instance, when the expert judge evaluates the methodology with which the entity develops its arguments.

Feigenbaum (2003) sets a further challenge for Expert Systems: their knowledge base should be independently acquired, rather than hand-coded by humans. The learning Expert System has two advantages for the FT: first, similarly to Turing's *child machine* (1948, 1950), it is intended to rule out the possibility for the system to be a pre-

[88] Feigenbaum (2003, p. 33).
[89] *Ibidem*.

programmed look-up table like Blockhead. Feigenbaum (2003) provides a description of the learning Expert System as follows:

> “First, manually encode a novice-level understanding (symbolic representation) of the domain, that is, humans will do the knowledge engineering. The novice-level “view” of the domain can be taken directly from a well-regarded elementary text of the domain, for example, an introduction to molecular biology. Second, write the software for a system that will read the “next” text in the field, augmenting as it reads the kernel novice-level knowledge base.” (37)

Second, the learning Expert System would not be left alone to learn new things: it would require some human assistance in order to occasionally cope with natural language, clarify some passage or correct the possible misunderstandings. And the humans assisting the Expert System could as well benefit from its discoveries. As Feigenbaum (2003) suggests:

> “The educated CI [Computer Intelligence] would continue to educate itself by reading the emerging literature of the domain. That is, it would “keep up with the literature.” Human assistance will still be allowed, but less than was allowed earlier. Indeed, one could think of this phase as “collaboration,” since both human and CI will be learning the new material at the same time.” (38)

To conclude, it is worth noting that the FT is compatible with Artificial Fallibility, but incompatible with Artificial Stupidity, for the entity has to show the same level of expertise in a certain domain as a human expert. In order to pass the FT, the Expert System has not only to show to possess the proper knowledge about a topic, but also to present and justify it in an expert human-like way.

Summing up, in this section I describe the subject matter expert test proposed by Feigenbaum (2003): the FT. I claim that it has the advantage, like the MIST, to prevent the entity from passing the test by successfully using Artificial Stupidity to exploit the judge's beliefs. But whereas the MIST evaluates the "quantity" of the subcognitive interrogation, the FT evaluates the "quality" of the expert conversation between the expert contestants. To do so, the FT parametrises the entity along the dimension of (i) 'expert human-likeness', that is, the ability to use a jargonized language properly enough to be recognised as a human expert; and the dimension of (ii) 'correctness', that is, the ability to produce the right responses to the expert questions and problems posed by the expert judge. Moreover, I argue, there is a further implicit dimension along which the entity is evaluated, that is, (iii) 'strategicness', which evaluates the methodology with which the entity produces and justifies its responses. The two main problems with the FT, however, are: (i) the FT is not a practical test, contrary to the TT; and (ii) the FT does not explicitly define the dimension of 'strategicness', leaving it to the judge's discretion.

4.3 The Extended Turing Tests

The expression "extended Turing Test" refers to any variation of the experimental design of the TT that includes an evaluation of other abilities, such as sensorimotor abilities, in addition to the conversational ones. In this section, I discuss the Total Turing Test (3T) proposed by Harnad (1989, 1991, 2000); and the Truly Total Turing Test (4T), proposed by Schweizer (1998, 2012a).

4.3.1 Total Turing Test

Harnad (2000) proposes a hierarchy of Turing tests, from the most basic to the most complex one:

> "from subtotal ("toy") fragments of our functions (t1), to total symbolic (pen-pal) function (T2 – the standard Turing Test), to total external sensorimotor (robotic) function (T3), to total internal microfunction (T4), to total indistinguishability in every empirically discernible respect (T5)." (425)

Harnad (2000) specifies that the models considered by the TT and its extended versions are models "of subhuman and human functional capacities;"[90] not animals or aliens. The reason is, Harnad holds, that even though "other species no doubt have minds, […] our confidence in this can only diminish as we move further and further from our own species."[91] In other words, without a general theory of intelligence, we can only focus on human-like intelligence, as held by French (1990).

Harnad (2000) argues that the t1 is underdetermined; the T2 is vulnerable to various difficulties; the T4 and the T5 are overdetermined; and the T3[92] is to be considered the proper experimental update to test for intelligence. More specifically, the t1 is the level where the evaluated models are arbitrary human functions only, and the candidates are more or less toys (consider, for instance, a talking doll or a chess engine). The T2 is the level where the scrutinised model is the human verbal ability, described as "words in – words out", or, better, "symbols in, symbols out"[93] (SISO). T4 and T5 are the levels

[90] Harnad (2000, p. 432).
[91] *Ibid*.
[92] It's worth noting that '3T' and 'T3' can be used interchangeably. On the contrary, '4T' and 'T4' cannot: the former refers to the Truly Total Turing Test (Schweizer, 1998) and the latter to the "Total Internal Microfunction" test, as distinguished by Harnad (2000).
[93] Harnad (1989), p. 19).

where the tested models are not only conversational ones but also neurophysiological ones: in the T4 the neurophysiological indistinguishability with humans is obtained through synthetic craftsmanship; in the T5 the neurophysiological indistinguishability with humans is engineered out of real biological molecules (for instance, an artificial embryo that will eventually grow into a human being). The T3, finally, is the level where the examined model is the external full human function-indistinguishability. In the T3 there is no interest to the inner workings or the nature of the inner components of a candidate, as long as the candidate is able to produce the full range of cognitive and physical behaviours of a human. As Harnad (2000) puts:

> "So does that mean that all differences above the level of T3 matter as little (or as much) to having a mind […]? Intuitively (and morally), I think the answer is undeniably: Yes." (441)

Summing up, the 3T is intended to extend the range of abilities that an entity must show, to include sensorimotor abilities in addition to conversational ones.

4.3.2 Truly Total Turing Test

Schweizer (2012b) agrees with Harnad about the weaknesses of the TT, but he is not satisfied with the 3T as the proper improvement of the TT's experimental design:

> "[…] I argue that the conversation-based 2T is far too weak, and we must scale up to the full linguistic and robotic standards of the Total Turing Test (3T). […] However, I then propose a variation on the 3T, adopting Dennett's

> method of 'heterophenomenology', to rigorously probe the robot's purported 'inner' qualitative experiences." (41)[94]

Whereas Harnad (1989, 2000) argues that the 3T represents the proper hierarchical level to which the TT should be updated, Schweizer (1998) objects that the 3T can only test for individual tokens of intelligent behaviour. This means that, even if the scrutinised candidate is embodied, there is no guarantee that it has experience of the world. The 3T robotic candidate could just translate the sensorimotor and conversational inputs its sensors receive into symbols, in order to manipulate them and emit the appropriate cognitive and physical pre-programmed outputs to interact with other agents and the environment. In other words, just as the TT can be passed by an unintelligent conversational Blockhead, the 3T can be passed by an unintelligent robotic Blockhead. Schweizer (1998) summarises the issue as follows:

> "It is not simply behavior that matters, but *how* the behavior is generated. For example, mere successful performance is not intelligent if it is produced accidentally, or if it is just parroting of acquired but uncomprehended phrases, or if it is produced non-constructively, as in the use of exhaustive look-up tables to find an answer in pre-packaged form." (265)

In agreement with Harnad (1989), who holds that the experimental design of the TT should not set any limit, Schweizer's (1998) proposal removes the restrictions of the TT:

[94] See Dennett (1992): "The heterophenomenological method neither challenges nor accepts as entirely true the assertions of subjects, but rather maintains a constructive and sympathetic neutrality, in the hopes of compiling a definitive description of the world according to the subjects." (83)

> "[…] to test for genuine intelligence, we would need to see how the computational procedures would perform over the long term, in a varied and unrestricted environment. Our behavioral interactions with fellow humans extend over protracted interval of time, and involve many kinds of activity other than mere verbal exchange." (265)

Those restrictions are: (i) the conversational input-output setup, (ii) the anonymity, and (iii) the time limit[95]. To recall them, (i) the conversational input-output setup has the advantage of keeping the test practical, but it cannot prevent the unintelligent Blockhead from passing. (ii) "Teletype anonymity is meant to screen-off prejudicial reactions, but the point then becomes to *fool* someone into thinking that the machine is human."[96] And (iii) the time limit is meant, as shown above, for practical reasons (an unlimited TT would not be passed by any mortal human as well). The reason to eliminate those restrictions is that "when it comes to judging human language users in normal contexts, we rely on a far richer domain of evidence."[97] Schweizer (1998) focuses on the "historical record"[98] of human intelligence, rather than on individual expressions of intelligence. The test is intended to evaluate whether a non-human cognitive type can replicate the human intellectual achievements. This "long-term evolutionary criterion"[99] is the dimension along which the entity, which is an individual token of the cognitive type under scrutiny, is parametrised in the Truly Total

[95] See (Harnad, 2000): "It is a mistake to think of T2 as something that can be "passed" in a single evening or even during an annual Loebner Prize Competition (Loebner, 1994; Shieber, 1994). Although it was introduced in the form of a party game in order to engage our intuitions, it is obvious that Turing intends T2 as a serious, long-term operational criterion for what would now be called "cognitive science." The successful candidate is not one that has fooled the judges in one session into thinking it could perform indistinguishably from a real pen-pal with a mind. (Fooling 70% of the people one time is as meaningless, scientifically, as fooling 100% of the people 70 times.) The winning candidate will really have the capacity to perform indistinguishably from a real pen-pal with a mind – for a lifetime, if need be, just as unhaltingly as any of the rest of us can." (433)
[96] Schweizer (1998, p. 266).
[97] Schweizer (2012a, p. 195).
[98] Schweizer (1998, p. 267).
[99] Schweizer (1998, p. 270).

Turing Test (4T)[100]. This perspective shift from individual cognitive tokens to general cognitive type is due to the nature of sociolinguistic competencies, as argued by Burge (1979)[101] and Schweizer (2012a):

> "Various *mental* states and events are not fully determined by what's going on in an *individual's head*. Instead, they rely in an inextricable manner on the encompassing sociolinguistic milieu. From this it follows that *human* mentality is essentially non-individualistic – it depends crucially upon a sociolinguistic context that transcends personal boundaries." (203)

Summing up, the 4T is an evolutionary evaluation of the capacities of a cognitive type (like humankind), rather than a circumscribed evaluation of the capacities of individual cognitive tokens (like humans). As, Schweizer (2012a) specifies, the point of the 4T is not to evaluate a cognitive token with a pre-existing sociolinguistic context, but rather "[…] whether the artificial cognitive type *itself* is capable of producing a comparable sociolinguistic medium."[102] The 4T is intended to show that not only the TT, but also the 3T is too weak[103]; and it is intended to face one in particular of the difficulties of the experimental design of the TT (and the 3T): the possibility for an unintelligent system which manipulates pre-programmed knowledge (like Blockhead) to pass. Also, it is important to clarify, in the 4T is not chauvinistic: the type's

---

[100] Schweizer's 4T (Truly Total Turing Test) is not to be confused with Harnad's T4 (Total Total Turing Test). See Schweizer (1998): "This [4T] is not to be confused with Harnad's '*TTTT*', which stands for 'Total Total Turing Test', and which includes neurophysiological imitation." (267)

[101] See Burge (1979): "I shall offer some considerations that stress social factors in descriptions of an individual's mental phenomena. These considerations call into question individualistic presuppositions of several traditional and modern treatments of mind. […] Even those propositional attitudes not infected by incomplete understanding depend for their content on social factors that are independent of the individual, asocially and non-intentionally described. For if the social environment had been appropriately different, the contents of those attitudes would have been different." (74-84)

[102] Schweizer (2012a, p. 191).

[103] See Schweizer (2012a): "my overall conclusion will be that the 3T is still too weak, and that a truly comprehensive test should evaluate the general *category* of cognitive organization under investigation, rather than the performance of single specimens." (198)

evolutionary history and the token's inner workings are not required to "be isomorphic to ours in order to count as intelligent"[104]. It is rather intended to show that the cognitive type under scrutiny should "be capable of analogous feats of adaptation and creation."[105]

4.4 Summary

In this chapter, I consider two alternative versions of the TT that provides some ground for my proposal: the MIST, proposed by McKinstry (1997, 2009); and the FT, proposed by Feigenbaum (2003). I also discuss two extended versions of the TT: the 3T, proposed by Harnad (1989, 2000); and the 4T, proposed by Schweizer (1998, 2000). The alternative versions are intended to deal with a specific problem of the TT. The extended versions are intended to include full sensorimotor abilities, in addition to verbal ones, among those that an entity should possess; and to extend the requirement for passing the test, from being indistinguishable from a human in holding a conversation to being indistinguishable from a human in doing everything a human can do.

The MIST is intended to evaluate the subcognitive competencies, as discussed by French (1990), of the candidate entity by means of a conversation carried out by yes/no interactions only. It has the advantage to prevent Artificial Stupidity from passing, but not Blockhead. The FT is intended to evaluate the expert competencies of the candidate entity by means of an expert jargonized dissertation. Like the MIST, it has the advantage to avoid Artificial Stupidity, but not Blockhead – or, at least, not all the times. The 3T is intended to evaluate the sensorimotor abilities of an entity (cognitive token), in addition to the conversational ones. The problem with the 3T is that it cannot prevent a robotic Blockhead from passing, just like the TT cannot avoid

[104] Schweizer (1998, p. 271).
[105] *Ibid.*

Blockhead. Finally, the 4T is intended to evaluate the evolutionary and sociolinguistic history of a cognitive type. No unintelligent approach could pass the 4T; however, it is a long-term, impractical test.

The extended versions of the TT are motivated by the view that the simple TT is not adequate and needs to be strengthened to embodied versions. In other words, according to these views, having direct human-like experience or a physical human-like body are necessary aspects along which the candidate entity should be parametrised. Even if I do not disagree with the advantages of an embodied agent, I keep the QTT a disembodied test in order to avoid chauvinistic consequences. For instance, the requirements of the extended TTs can be rejected by the following counterexample: an individual, born with serious sensorimotor deficits and incapable of directly experiencing the world, who is still able to fully develop human cognitive abilities. However, as mentioned in the Conclusion, future work will focus on extended versions of the QTT.

In the next chapter, I finally present my new test: the QTT.

Chapter 5

The Questioning Turing Test

Abstract. In this chapter, I describe the new version of the TT that I propose: the Questioning Turing Test (QTT), where the conversation between the participants is an enquiry rather than an open one. The QTT is made of two procedures: (i) a questioning game played by two humans and (ii) a questioning game played by a human and a machine. The two games have a similar scope as the AB-IG and the MB-IG (Ch. 1), that is, to set a benchmark in order to compare the two outcomes, where the former scores the latter. Moreover, the roles of the questioner and the answerer are switched: here, the candidate entity asks the questions and the judge gives the replies. The entity is parametrised along two further dimensions in addition to 'human-likeness': 'correctness' and 'strategicness'. This, I claim, prevents both Artificial Stupidity (Ch. 2) and Blockhead (Ch. 3) from passing.

In this chapter, I propose my new version of the TT: the Questioning Turing Test (QTT). The QTT focus on a specific kind of conversation, that is, enquiries. It's important to point out that the TT already contains every possible conversational variation of the test, and it follows that the QTT is already included in the TT. In other words, nothing prevents the conversation during the TT to be an enquiry. However, I disagree with the view that the TT already includes any test, as argued by Hernández-Orallo (2017):

> "[…] the Turing test can 'contain' any test we may think of, so the question backtracks to what questions the judge must ask to make it sufficient." (p. 129)

The reason is that, while the open conversation in the TT admits any possible conversational variation, it does not necessarily follow that the experimental design of the original TT also includes any possible experimental variation. If the judge purposely runs a variation of the TT, then the judge and the experimenter are the same person; and if that were the case, the judge should be regarded as an *expert*. This contradicts what Turing specifies about the judges, "who should not be expert about machines[106]", excluding them from participating (as well as "those who are expert about the human mind[107]"). Turing, on the contrary, specifies that the judge should rather be "an average interrogator[108]", where the function of *average* is to rule out those "conditions that might generate false positives or negatives – or whatever stands proxy for a false result […]."[109] Moreover, the statement of Hernández-Orallo implies that the judge can objectively evaluate the entity's intelligence alone, no comparisons with any benchmark needed, given that the judge asks the right questions. This contradicts the "Literal Interpretation" of the TT (discussed in Ch. 1), according to which the results are given by the comparison between the contestants' performances in the AB-IG and the MB-IG. So, I hold that it is true that the TT can *virtually* contain any variation, for it is, as Traiger (2000) states, best regarded as a format to test for intelligence which "invites generalization"[110] and potential modifications; but I also believe that, because of this, the TT is exploitable, and its reliability should not depend upon the human judge. My claim is that it is necessary to improve the experimental design of the TT to minimise both the Eliza Effect and the Confederate Effect (Ch. 1); and to avoid both Artificial Stupidity (Ch. 2) and Blockhead (Ch. 3).

The chapter is structured as follows: in the next section, I show the switch from SISO to SOSI. In section 5.2, I describe the QTT. In section 5.3, I outline the main differences between the TT, the MIST, the FT and the QTT; and I show the advantages of the QTT over the TT, the

[106] Turing (1951, p. 495).
[107] Copeland (2000, p. 525).
[108] Turing (1950, p. 442).
[109] Proudfoot (2013, p. 397).
[110] Traiger (2000, p. 565).

MIST and the FT. In section 5.4, I show the Hybrid QTT (where the entity is played by both a human and a machine). In section 5.5, I discuss the potential extended versions of the QTT.

5.1 From SISO to SOSI

The new test that I propose focuses on two aspects. First, the comparison between the entities' performances. Like Sterrett's OIG (2000), the QTT evaluates:

> "[…] whether the machine is as resourceful in using *its* resources in performing a difficult task as the man is in using *his* resources in performing the same difficult task." (548)

The comparison between the resourcefulness of the entities in performing the same task provides the benchmark with which their performances are scored.

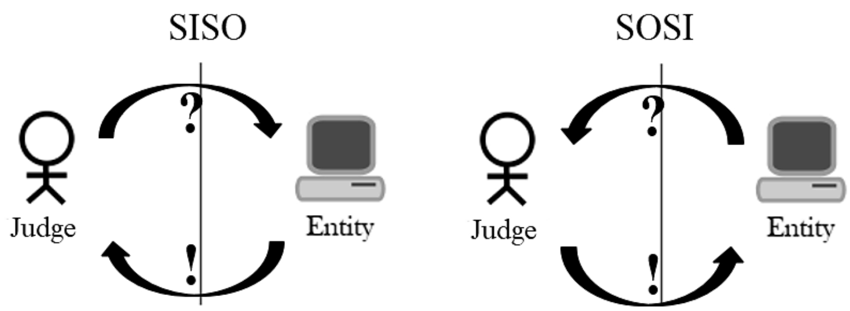
Second, the QTT is designed to evaluate *how* an entity's responses or behaviours are produced. I agree with Schweizer (1998) that:

> "[…] it is the method or procedure for generating behavior which is crucial to the issue, and that behavior itself is important mainly insofar as it provides data for gauging the respective quality of these underlying methods. In this vein, it is frequently argued that the most serious *defect* of the original *TT* is precisely the fact that the method by which the behaviour is produced is not adequately probed. […] Intelligence is not just a question of isolated samples of behavior, but of *how* the behavior is produced." (265-268)

However, in a SISO test, like the TT, I hold that it is not possible to evaluate how a behaviour is produced. The reason is that, in a SISO test, the interactions are unidirectional: from the inputs of the human judge to the outputs of the candidate entity, and back again. As McKinstry (2009) puts it, the entities:

> "[…] are usually stimulus-response only. That is, they are able to respond in a perfectly human-like fashion to previously anticipated stimuli and an approximately human-like fashion to unanticipated stimuli, but they are incapable of generating original stimuli themselves." (p. 296)

As introduced in the previous chapter, the setup of the TT can be defined as "symbols in, symbols out"[111] (SISO). The setup of the QTT, on the other hand, is switched to "symbols out, symbols in" (SOSI). In other words, the candidate entity asks the questions to accomplish an enquiry, and the human judge provides the responses (see fig. 1).



(fig. 1)

The QTT switches the experimental structure from SISO (symbols in, symbols out), where the entity emits an appropriate response for a given

[111] Harnad (2000, p. 443).

input, to SOSI (symbols out, symbols in), where the entity produces an appropriate input for a given output. In other words, whereas in a SISO test, like the TT, the entity replies to the judge's questions, in a SOSI test, like the QTT, the entity asks questions and the judge replies.

An important characteristic of the experimental design of the TT is that, since it focuses on the entity's replies, the entity only needs to manipulate the symbolic inputs (the judge's questions) in order to send symbolic outputs. Blockhead, for instance, receives the first interaction from the judge, and then it proceeds in looking its table for all the compatible entries and picks an *appropriate* one to send. If Blockhead were to produce some interaction first, it would just be a randomly picked one among the entries in its table (and, therefore, unlikely an appropriate one). French (2000a) considers SISO tests unreliable ones, since, as he states:

> "Mere SISO questioning could allow a non-intelligent entity to slip through the cracks of the Turing Test." (334)

The SOSI setup of the QTT allows the entity to be parametrised along a further dimension in addition to 'human-likeness' and 'correctness': 'strategicness', which refers to the number of questions asked by the entity in order to accomplish the aim of the enquiry (or before giving up): the fewer the questions asked, the better the strategy used. So, in the QTT the questioning entity is evaluated in terms of:

> (i) 'human-likeness', that is, the ability to be recognised as human by the judge as frequently as a human;
> (ii) 'correctness', that is, the ability to accomplish the aim of the enquiry;

(iii) 'strategicness', that is, the ability to ask as few – human-like – questions as possible in order to accomplish the aim of the enquiry in an optimised way.

I hold that the QTT avoids chauvinism, by parametrising the entity along 'human-likeness', as well as Artificial Stupidity, by parametrising the entity along 'correctness', and Blockhead, by parametrising the entity along 'strategicness'.

5.2 The Questioning Turing Test

Following Sterrett's OIG, the QTT is made of two procedures: (i) the human-questioning-human (HqH) and (ii) the machine-questioning-human (MqH). The contestants, both in the HqH and in the MqH, are in separate rooms communicating through text-based interactions. Depending on the aim of the enquiry, the QTT can be adapted to be played by either an *average* or an *expert* answerer. For instance, if the candidate entity is intended to carry out an enquiry about molecular biology, the answerer should be a molecular biology expert; if the candidate entity is intended to carry out a personal enquiry about the answerer (e.g. medical anamnesis) the answerer can be an average person.

In what follows, I describe the two procedures involved in the QTT: the HqH and the MqH. The first procedure of the QTT is the HqH, where only humans are involved. The candidate entities have the goal to strategically accomplish the aim of the enquiry and the judge has the goal to decide over their 'human-likeness'. The HqH has the same purpose as the AB-IG in the OIG, that is, to provide a benchmark with which the second procedure is scored. The second procedure of the QTT is the MqH, where one of the human questioners is replaced by a questioning machine. Similarly, in the MqH the candidate entities have the goal to accomplish the aim of the enquiry, and the judge has the goal to decide over their ‘human-likeness’.

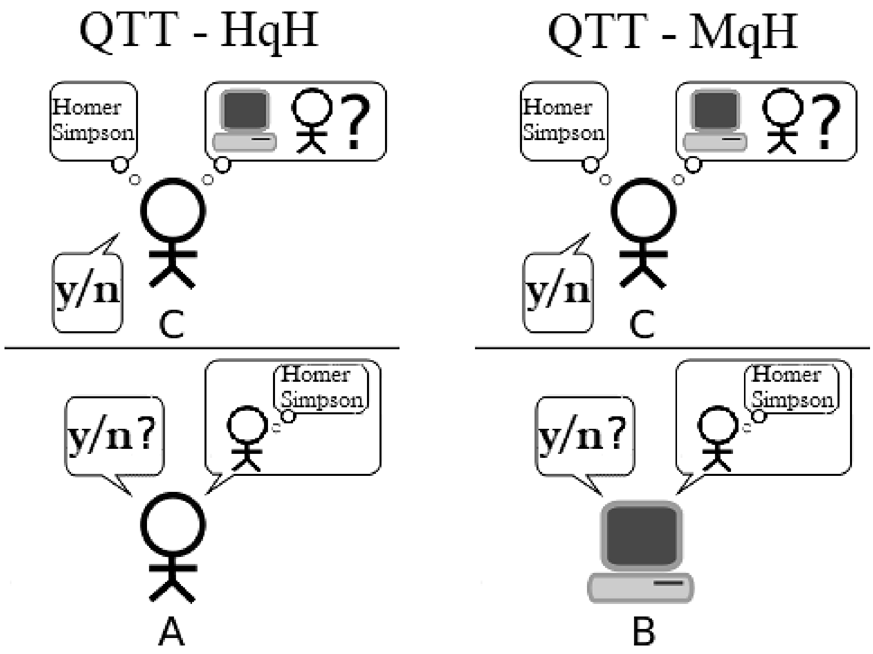
I distinguish between two kind of QTT. On the one hand, the open, unrestricted QTT is intended to be played *parallel-paired*, where there are two rounds in total, and each round involves 3 players: (A) a questioning human/machine, (B) a questioning human and (C) the answering judge. On the other hand, the practical version of the QTT follows the third experimental setup provided by Turing: the one-to-one test, between A (candidate entity) and C (the judge):

> “The [imitation] game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether someone really understands something or has “learnt it parrot fashion.” (p. 446)

The reasons why the practical, yes/no QTT is conducted *viva-voce* rather than *parallel-paired*, are: (i) the game involves a cooperative enquiry, rather than a competitive impersonation like the TT or the unrestricted QTT (where a parallel comparison between the performances of the entities is needed); and (ii) the language is restricted to yes/no interactions, and the importance of the interactions’ style is reduced. The practical QTT involves a twenty-questions game between two contestants with no particular expertise required. The answerer thinks about a public figure and the questioner has to identify who is being thought of, by gaining information from the answerer by means of yes/no questions. Like in the unrestricted QTT, the candidate entity in the practical QTT is parametrised along three dimensions: ‘human-likeness’, which evaluates its ability to be considered human thanks to the style of the questions it asks; ‘correctness’, which evaluates its ability to identify the public figure; and ‘strategicness’, which evaluates its performance in terms of the number of questions needed to accomplish the enquiry – the fewer the better. So, in the practical QTT, the entity passes if: (i) it produces human-like questions, that is, questions that the answerer recognises as human as frequently as those asked by a human; (ii) it successfully identifies the public figure as often as a human does; and (iii) it shows a strategic method to carry out the enquiry, that is, the number of questions it needs to accomplish the enquiry does not exceed the number of questions needed by a human questioner (see fig. 2).

In the HqH of the practical QTT, the goal of the human questioner is to ask strategic yes/no questions (where ‘strategic’ means as few as possible) in order to gain the information needed to successfully identify the public figure the answerer is thinking of. In the MqH of the practical QTT, the goal of the questioning machine is to ask strategic yes/no questions (where ‘strategic’ means as few as possible) in order to gain the information needed to identify the public figure the answerer is thinking of. The machine passes the test if it is recognised as human with – at least – the same frequency with which the human questioner is recognised as human in the HqH; and if it successfully identifies the public figure the judge is thinking of with no more questions than the human questioner in the HqH. In other words, the machine passes the MqH if its performance matches or exceeds the human performance in the HqH. This is consistent with Sterrett’s view that the machine, to be attributed with intelligence, has to be “sufficiently resourceful at a difficult task […] to win more rounds than the man or men playing the game do.”[112]

[112] Sterrett (2000, p. 548).



(fig. 2)[113]

As follows, I show a yes/no enquiry in the practical QTT, where the entity tries to identify the public figure (Alan Turing) that the judge is thinking of.

Entity: “Is your character real?”
Judge: “Yes.”
Entity: “Is your character a female?”
Judge: “No.”
Entity: “Is your character British?”
Judge: “Yes.”
Entity: “Is your character a youtuber?”
Judge: “No.”
Entity: “Is your character linked with sports?”
Judge: “No.”
Entity: “Is your character older than 18?”
Judge: “Yes.”

[113] In (fig. 2) I show the HqH and the MqH of the practical version of the QTT, where the questioning hidden entity (A or B) has to identify, by means of no more than 20 binary questions, the public figure the answerer (C) is thinking of. In the HqH, C thinks of a public figure (in this case Homer Simpson) and A asks yes/no questions to try to guess it; at the end, A states what C is thinking of (in this case A correctly identifies the public figure as Homer Simpson) and C decides whether A is human or machine. In the MqH, C thinks of a public figure and B asks yes/no questions to try to guess it; at the end, B states what C is thinking of and C decides whether B is human or machine.

Entity: "Is your character more than 40 years old?"
Judge: "Yes."
Entity: "Is your character still working?"
Judge: "No."
Entity: "Has your character been dead for more than 20 years?"
Judge: "Yes."
Entity: "Has your character been dead for more than 100 years?"
Judge: "No."
Entity: "Did your character take part in World War II?"
Judge: "Yes."
Entity: "Is your character a politician?"
Judge: "No."
Entity: "Has your character killed humans?"
Judge: "No."
Entity: "Is your character homosexual?"
Judge: "Yes."
Entity: "Does your character owe his fame to computing?"
Judge: "Yes."
Entity: "Does your character solve puzzles?"
Judge: "Yes."
Entity: "Is your character suicidal?"
Judge: "Yes."
Entity: "I think of Alan Mathison Turing."

5.3 Differences between the QTT and the TT/MIST/FT

I argue that there are three main differences between the TT/MIST/FT and the QTT.

(i) First, in the TT, the MIST and the FT, the entity must prove it is able to emit adequate responses to a series of inputs, where the inputs are, respectively, lines of a general conversation, subcognitive questions, and lines of an expert conversation. In the QTT, in contrast, the entity must prove it is able to generate adequate outputs in order to receive the inputs it needs to accomplish its task, where the outputs are strategic questions, the inputs are the answers received, and the task is to accomplish an enquiry. The questions are strategic when the machine is able to accomplish the enquiry by asking fewer questions than, or as many questions as, the human questioner. It is worth noting that it's not necessary for the machine to ask the same questions asked by the human (since different paths can lead to the same result), just that the number of questions asked does not exceed the number of questions asked by the human. This switch in terms of the task corresponds to the switch in terms of the setup of the test: whereas the TT is a SISO test, the QTT is a SOSI one.

(ii) Second, the human judge in the QTT plays a different – and less troubled – role than the judge in the TT and the FT (in the MIST, the judge can be entirely replaced with an automatized system, since the hidden entity has to reply to previously stored yes/no subcognitive questions). The judge in the QTT, on the one hand, can only interact during the test by means of yes/no answers; the judges in the TT and the FT, on the other hand, can interact during the test by means of, respectively, any conversational style whatsoever and an expert conversational style. As a result, in the TT (and, to a lesser extent, the FT), the judge may have various biases that cause her to engage the candidates in rather odd conversations in order to unmask them, making the test dependent on the judge's beliefs. This is especially true for the STT, as pointed out by Traiger (2000), where the judge will not talk in the same way she would talk with other humans. The judge will rather engage the hidden entity in a challenging and odd conversation, made of paradoxical questions and semantic traps in order to unmask a potential machine. In order to avoid this problem, in the QTT the human judge (that is, the answerer) cannot alter the conversation in any way and, thus, she cannot alter the test with personal beliefs or biases, as it can potentially happen in the TT and the FT. A further advantage of the QTT is that the role of the judge is more versatile: whereas the judge in the TT is required to be average, and the one in the FT is required to be expert, the judge in the QTT may be either average or expert (depending on the aim of the enquiry), with no risk of biased or chauvinistic consequences.

(iii) Third, the experimental design of the QTT parametrises the candidate entity along three dimensions: 'human-likeness', 'correctness' and 'strategicness'. It means that the benchmark scoring the test is not only based, as it is in the original TT, on 'human-likeness', which evaluates the frequency with which the human judge decides wrongly about the nature of the contestants; and it is not mainly based, as it is in the MIST and the FT, on 'correctness', that is, the entity's ability to reply correctly to, respectively, subcognitive and expert questions. I claim that, considered singularly, these two parameters lead to biased results: 'human-likeness' alone could be achieved by an unintelligent chatbot, by means of strategies such as Artificial Stupidity; and 'correctness' alone could be achieved an Expert System which shows no 'human-likeness' whatsoever (like a calculator). I rather hold we should evaluate both 'human-likeness' and 'correctness', in order to avoid both Artificial Infallibility and Artificial Stupidity. However, evaluating both 'human-likeness' and 'correctness' is still not enough, for Blockhead could pass by producing the appropriate reply by means of a brute-force, non-strategic approach and a huge look-up table. In order to prevent this from occurring, I propose to parametrise the candidate entity along 'strategicness', intended to evaluate *how well* the entity can accomplish a task. In the practical QTT, where the candidate entity has to accomplish the aim of the enquiry by means of yes/no questions, 'strategicness' is defined in terms of the number of questions asked, the fewer, the better. A questioning Blockhead would be prevented from passing because it would either search through its database one question at a time, potentially taking an incredibly long time to accomplish the enquiry; or it would fail to ask as few questions as possible, without an algorithm to optimise the search. ‘Strategicness’, I claim, can only be evaluated in a SOSI test, as the QTT, and not in a SISO one, where a strategic entity could never be discriminated from an unintelligent system like Blockhead.

Summing up, in the (SISO) TT the hidden entity interacts with an average human judge, and its goal is to be recognised as a real human. In the (SISO) MIST, the hidden entity interacts with a human or an automated system to demonstrate to be able to answer 20 yes/no subcognitive questions correctly, in order to provide statistical evidence of its intelligence. In the (SISO) FT, the hidden entity interacts with a human expert, in order to be recognised as a human expert. In the (SOSI) QTT, the human judge can be either average or expert, and the goal of the candidate entity is to (i) be recognised as human, (ii) accomplish the aim of the enquiry (avoiding Artificial Stupidity) and (iii) do so with as few questions as possible (ruling out Blockhead). The following table (tab. 1) is intended to illustrate the differences between the TT, the MIST, the FT and the QTT.

Test	**Parameters**	**Setup**	**Judge**	**A. S.**	**Blockhead**
TT	Human-likeness	SISO	Average	Pass	Pass
MIST	Correctness	SISO	Trained	Fail	Pass
FT	Expert-likeness Correctness	SISO	Expert	Fail	Pass
QTT	Human-likeness Correctness Strategicness	SOSI	Average or Expert	Fail	Fail

(tab. 1)[114]

[114] A. S. stands for Artificial Stupidity.

5.4 Hybrid QTT

The experimental design of the QTT, I argue, has a further experimental advantage over the TT, the MIST and the FT. The QTT can be used to test Hybrid Systems. A Hybrid System can be described as a system in which a human and a machine work together, combining their strengths and abilities, performing better than by themselves considered individually. Hybrid Systems has gained a growing interest in recent years, for instance, Sinha & al. (2016) write:

> "Current machine algorithms for analysis of unstructured data typically show low accuracies due to the need for human-like intelligence. Conversely, though humans are much better than machine algorithms on analysing unstructured data, they are unpredictable, slower and can be erroneous or even malicious as computing agents. Therefore, a hybrid platform that can intelligently orchestrate machine and human computing resources would potentially be capable of providing significantly better benefits compared to either type of computing agent in isolation." (1)

And Demartini (2015) shows (see tab. 2) the most recent developments of hybrid systems in the last decade, and he points out:

> "The creation of hybrid human-machine systems is a highly promising direction as it allows leveraging both the scalability of machines over large amounts of data as well as keeping the quality of human intelligence in the loop to finally obtain both efficiency and effectiveness in data processing applications." (5)

Year	Domain	Data type	Human role	Incentive	Time constrains
2006	Web	Images	Pre-p.	Fun	Batch
2007	Science	Images	Pre-p.	Community	Batch
2008	Web	Images	Post-p.	Access	Batch
2011	Database	Graph	Pre-p.	Monetary	Batch
2011	Database	Struct. data	Pre-p.	Monetary	Real-time
2011	Filtering	Video	Pre-p.	Monetary	Real-time
2012	Database	Struct. data	Post-p.	Monetary	Real-time
2012	Web	Unstruct. text	Post-p.	Monetary	Batch
2012	Data integration	Struct. data	Post-p.	Monetary	Batch
2012	Entity resolution	Struct. data	Post-p.	Monetary	Batch
2012	Entity resolution	Struct. data	Post-p.	Monetary	Batch
2012	Search	Unstruct. text	Post-p.	Community	Real-time
2012	Captioning	Video	Pre-p.	Community	Real-time
2013	Info extraction	Unstruct. text	Post-p.	Monetary	Batch
2013	Entity resolution	Struct. data	Post-p.	Monetary	Batch
2013	Entity resolution	Struct. data	Post-p.	Monetary	Batch
2013	Database	Struct. data	Pre-p.	Monetary	Batch
2013	Database	Struct. data	Post-p.	Monetary	Real-time
2013	Biomedical	Ontology	Pre-p.	Monetary	Batch
2013	Personal assistance	Unstruct. text	Pre-p.	Monetary	Real-time
2013	Biomedical	Unstruct. text	Post-p.	Fun	Batch
2014	Search	Image	Pre-p.	Monetary	Real-time
2014	Database	Struct. data	Post-p.	Monetary	Real-time
2014	Cult. heritage	Image	Pre-p.	Monetary	Batch

(tab. 2)[115]

As (tab. 2) shows, Hybrid Systems have great potential in many domains. However, at present, the human component of Hybrid Systems plays a marginal role, giving a contribution only before or after the machine's performance, and not in real-time, as the Hybrid QTT would require. In the next section, I provide three examples of unrestricted QTTs and Hybrid QTTs.

5.5 Unrestricted Versions of the QTT

As mentions above, the reasons why the practical QTT is restricted to yes/no enquiries is to conduct as many tests as possible for my experiment, as quickly as possible. It is, however, not to a mandatory restriction. In this section, I present three alternative and unrestricted

[115] Demartini (2005): "The columns indicate respectively the year of publication, the domain of application of the hybrid human-machine system, the type of data processed by the system, the role of the human component in the hybrid human-machine system (i.e, processing data either before or after the machine component), the type of incentive used to motivate crowd workers to perform tasks, and finally the time constrains of the hybrid system, that is, whether the human-machine system performs batch or real-time data processing." (8)

versions of the QTT that would require full use of natural language, and where 'human-likeness', 'correctness' and 'strategicness' would play a much more evident role than they do in the practical QTT.

(i) *Detective QTT* – The Detective QTT[116] is played *parallel-paired* between the human Detective (A), the machine Detective (B) and the Suspect (C). A and B have to ask strategic (meaning the fewer, the better) questions in order to provide the right verdict (C is guilty or C is innocent); and C, in turn, has to decide who the human detective is. The idea is that questioning programs might become handy to security, and they could also be implemented in Hybrid Systems, where the interrogations are cooperatively (rather than competitively) carried by both the machine and a human detective.

(ii) *Medical QTT* – The Medical QTT is played *parallel-paired* between the human Doctor (A), the machine Doctor (B) and the Patient (C). A and B have to ask strategic questions in order to accomplish an enquiry about the medical condition of C who, in turn, has to decide who the human doctor is. The idea is that questioning programs might become the first point of contact for primary health care, and they could also be implemented in Hybrid Systems, where each case is cooperatively examined by a human doctor and a medical machine.

(iii) *Customer Service QTT* – The Customer Service QTT is played *parallel-paired* between the human Operator (A), the machine Operator (B) and the Customer (C). A and B have to ask strategic questions in order to properly assist C, who decides who the human Operator is. The idea is that questioning programs might become the first point of contact for customer service, and they could also be implemented in Hybrid Systems, where each query is cooperatively handled by a human operator and a machine.

[116] A good example could be something like *Blade Runner*'s Voight-Kampff test.

5.6 Summary

In this chapter, I describe the QTT in order to improve the experimental design of the TT and avoid Artificial Stupidity and Blockhead. The main novelty in the QTT is the switch from the SISO setup (as in the TT, MIST and FT) to the SOSI setup, where the role of the judge and the candidate entity are reversed: the latter asks the questions and the former provides the replies. This also gives a further advantage for the choice of the human judge, who can be either an average person or an expert, depending on the aim of the enquiry (whereas, it's worth recalling, in the TT the judge has to be an average person, in the FT an expert, and in the MIST a trained one). The SOSI setup – and, by extension, the questioning process – allows to parametrise the hidden entity along a further dimension, in addition to 'human-likeness' and 'correctness': that is, 'strategicness', intended to evaluate the method used by the candidate in terms of the number of questions needed to accomplish the enquiry. So, in order to pass the QTT, a candidate entity must (i) ask human-like questions, (ii) accomplish the aim of the enquiry and (iii) show a good method by asking as few questions as possible.

The yes/no QTT, that I run for practical reasons (that is, to make the tests as quick and simple as possible), is conducted *viva voce*. This is because the game is a cooperative enquiry, rather than a competitive impersonation like in the TT. The unrestricted QTT, on the other hand, is played *parallel-paired*, not *viva-voce*. However, it is worth noting, the *viva-voce* QTT involves two procedures: a human-questioning-human test (HqH) and a machine-questioning human test (MqH). The results, similarly to the OIG described by Sterrett (2000), are given by the comparison between the machine's performance in accomplishing a difficult task – which requires resourcefulness – and the human's performance in accomplishing the same task. A further advantage of the experimental design of the QTT over the TT and its alternative versions, is that it allows a hybrid version to be run, where the role of the hidden entity is played by both the machine and the human. The Hybrid QTT is intended to show that the performance of the hybrid human-machine system is better compared to the performances of the two systems alone. Finally, I point out that the experimental design of the QTT does not require the verbal interactions to be restricted to binary ones, as in its practical version. A number of different variations of the QTT and Hybrid QTT can be designed without this verbal restriction. These versions are *parallel-paired* ones, such as the Detective QTT (where a human detective and a machine detective have to interrogate a suspect); the Medical QTT (where a human doctor and a machine doctor have to take the medical history of a patient); or the Customer Service QTT (where a human operator and a machine operator have to assist a customer).

Chapter 6
The Objections

Abstract. In this chapter I discuss three difficulties that can be raised against the QTT: (i) the first claims that the QTT is redundant; (ii) the second claims that the QTT is chauvinistic; (iii) the third claims that the QTT cannot prevent Blockhead from passing. To reply, I argue that (i) the QTT is not redundant, because the experimental variation of the QTT is not included in the TT; (ii) the QTT is not chauvinistic, because a human-like, incorrect and non-strategic approach could still pass the test; (iii) Blockhead, as well as Expert, Stupid and Ultimate Blockhead, cannot pass without an information-gathering algorithm to optimise (and speed up) their processes.

This chapter is dedicated to the objections to the QTT. I consider three main objections to the QTT: the first one, discussed in the next section, claims that the QTT is redundant, for it is already included in the TT. The second one, discussed in section 6.2, claims that the QTT is chauvinistic, for an intelligent agent could still fail to strategically accomplish an enquiry. The third one, discussed in section 6.3, claims that the QTT cannot prevent Blockhead from passing, for it is logically possible to build a general questioning Blockhead, and it is even physically feasible to build an expert questioning Blockhead.

6.1 The QTT is Redundant

The first objection I consider is that the QTT is redundant, for the TT "can contain any test we may think of"[117], QTT included.

[117] Hernandez-Orallo (2017, p. 129).

I reject this view. Or, better, I admit that the TT includes any possible *conversational variation*: that is, in the TT the judge can decide to engage any kind of conversation (dialogues, monologues, dissertations, interrogations, and so on). However, it does not follow that the TT also includes any possible *experimental variation* of the TT. The reason is the following. If the judge purposely runs an experimental variation of the TT, then the judge is also the experimenter and, implicitly, an expert of some sort. And this contradicts what Turing specifies about the judges, who should rather be "an average interrogator[118]", where the function of 'average' is to rule out false positives or negatives. The judge, in other words, should not be an expert of some sort and, by extension, an experimenter. That said, I do believe that the TT can *virtually* contain any experimental variation, for it is best regarded as a format to test for intelligence which "invites generalization"[119] and potential modifications. In other words, I hold that the experimental design of the TT is intended to be adaptable to different experimental variations, but it is not for the – average – judge to design or run such experimental variations.

6.2 The QTT is Chauvinistic

The second objection that I consider is that the QTT is chauvinistic. The argument can be summed up as follows:

> By failing to perform in terms of 'correctness' and 'strategicness', that is, by failing to accomplish the aim of the enquiry by means of an optimised process (i.e. with as few questions as possible), intelligent agents who take the QTT may be considered unintelligent.

[118] Turing (1950, p. 442).
[119] Traiger (2000, p. 565).

I reply to this objection by clarifying that to pass the QTT the entity still needs to be recognised as a human. As I show in Ch. 9, where I discuss the results of my study, entities that fail in terms of ‘correctness’ and ‘strategicness’, but succeed in terms of ‘human-likeness’, are still plausible candidates to be attributed with intelligence. Conversely, entities that succeed in terms of ‘correctness’ and ‘strategicness’, but fail in terms of ‘human-likeness’, fail the QTT. ‘Human-likeness’, just like in the TT, is the parameter intended to justify the entity being attributed with intelligence in the first place. ‘Correctness’ and ‘strategicness’, on the other hand, are intended to prevent, respectively, Artificial Stupidity and Blockhead from exploiting the judge’s notion of ‘human-likeness’. In other words, in the QTT the attribution of ‘human-likeness’ is supported by the evaluation of ‘correctness’ and ‘strategicness’, whereas in the TT the attribution of ‘human-likeness’ is supported by no further evaluations other than the judge’s decision.

I rather hold that the QTT is less chauvinistic than the TT, MIST and FT. The TT requires the judge to be an average person, implicitly banning, for instance, individuals with certain conditions. The MIST requires the judge to be trained in subcognitive games. The FT requires the judge to be an expert in a certain domain. In contrast, the judge in the QTT can be either an average person or an expert; and, in both cases, the QTT is still able to avoid both Artificial Stupidity and Blockhead.

6.3 The QTT Cannot Avoid Blockhead

The third objection I consider is that the experimental design of the QTT is not able to prevent Blockhead from passing. According to this, it would always be possible for the questioning Blockhead[120] to produce the appropriate sequence of questions in the QTT, just like the conversational Blockhead can always produce the appropriate sequence of responses in the TT.

To recall Blockhead (Ch. 3), it is a system designed to look up its gigantic – but finite – table of interactions and to pick a sensible response to any verbal input whatsoever. It is worth noting that Blockhead is designed to *reply* to the interlocutor's verbal interactions, not to *produce* new ones. This experimental setup, focused on the entity's conversational outputs, is called "symbols in, symbols out" (SISO[121]), and I hold that a SISO test cannot avoid Blockhead.

My argument is that (i) the switch from SISO to SOSI in the QTT allows to measure 'strategicness' in addition to 'human-likeness' and 'correctness'; and (ii) Blockhead would never be able to perform well enough in terms of 'human-likeness', 'correctness' and 'strategicness' altogether. I show that the QTT is able to avoid not only the full LUT (Blockhead), but also the other versions discussed in Ch. 3: the specific LUT (Expert Blockhead), the uncooperative LUT (Stupid Blockhead) and the learning LUT (Ultimate Blockhead).

(i) Block suggests that Blockhead can be attributed with intelligence *if and only if* there is some upper bound to the *duration* of its testing. In the QTT, while I do not reject the duration limit (an entity that takes hours to play the QTT would not pass, failing to be attributed with 'human-likeness'), I propose a different kind of restriction: the number of interactions, that is, the number of questions asked ('strategicness'). Now, it may seem that such a restriction would facilitate Blockhead from passing the QTT, since it would need to pick just a few questions

[120] In this section, every time I discuss Blockhead in the context of the QTT I implicitly refer to the questioning Blockhead.
[121] See Harnad (1989).

from its table. The point is that those questions need to be strategically selected, in order to maximise the information acquisition needed to accomplish the aim of the enquiry.

(ii) There are at least four versions of Blockhead that it's worth discussing: (i) the full LUT (Blockhead); (ii) the small, specific LUT (Expert Blockhead); (iii) the uncooperative Elizish LUT (Stupid Blockhead); and (iv) the independently acquired LUT (Ultimate Blockhead). As follows, I consider each case individually, and I discuss how each version of Blockhead would do in both the *practical* and the *unrestricted* QTT. Recalling the distinction, in the *practical* QTT the hidden entity has to identify the public figure being thought of by the human judge, by asking as few yes/no human-like questions as possible; while in the *unrestricted* QTT, the hidden entity has to accomplish the aim of an open enquiry, by asking as few human-like questions as possible.

6.3.1 Full LUT (Blockhead)

I argue that there are two reasons why the full LUT (Blockhead) can pass neither the *practical* nor the *unrestricted* QTT. The first is that Blockhead is potentially very slow[122]. Blockhead can only search its table one *item* (that is, *response*, in the TT; *question*, in the QTT) at the time[123]. The second reason is that Blockhead would not be able to accomplish the aim of an enquiry by asking as few questions as possible, since it lacks an algorithm to optimise the search through its table. Instead, Blockhead would ask random and pointless questions. The importance of such an algorithm is advocated by Russell & Norvig (2010), who claim that information-gathering systems must be able to acknowledge the value of information (see Ch. 7).

[122] See Copeland (2000): "what the brain can do in minutes would take this machine thousands of millions of years." (533)
[123] See Block (1981).

So, in the *practical* QTT, Blockhead would be able to produce human-like yes/no questions (for its full table would contain any human-like yes/no question) quick enough (since the enquiry is a specific, yes/no one); and it would also be able to accomplish the aim of the yes/no enquiry. However, Blockhead would fail to do so with as few questions as possible, since it lacks an information-gathering algorithm to optimise the search through its huge table of questions. Blockhead would rather ask random questions, failing thus not only ‘strategicness’, but also ‘human-likeness’. On the other hand, in the *unrestricted* QTT, Blockhead would take a very long time to ask a question, due to the size of its table. So, even if it would be able to ask questions in a human-like style, it would not be attributed with ‘human-likeness’, due to its slowness in asking them. Moreover, even though Blockhead would eventually be able to accomplish the aim of any enquiry, it would fail, as in the *practical* QTT, to do so with as few questions as possible (being equipped with no information-gathering algorithm to optimise the search through its table), failing thus ‘strategicness’.

6.3.2 Specific LUT (Expert Blockhead)

In case the enquiry is a specific one, and Blockhead’s table is small enough, Blockhead would be able to show ‘strategicness’ even without an algorithm to optimise its search. Even so, I hold that the small and specific LUT (which I call Expert Blockhead) can pass neither the *practical* nor the *unrestricted* QTT. Expert Blockhead’s table contains certain verbal interactions only (in the case of the *practical* QTT, yes/no questions) with a specific purpose (in the case of the *practical* QTT, to identify the public figure that the human judge thinks of).

In the *practical* QTT, where the enquiry is not an open one and the interactions are limited to twenty yes/no questions and replies, Expert Blockhead would not take so long to search its smaller, specific LUT to accomplish the aim of the enquiry correctly. And it would even be able to do so with very few questions. However, the style, that is, the 'human-likeness' of Expert Blockhead's questions would be sacrificed (the table would be too small to include every human-like yes/no question). In other words, although Expert Blockhead would be able to accomplish the aim of the enquiry with a very few questions, it would fail the *practical* QTT by not being attributed with 'human-likeness', due to the style of its questions. In Ch. 9, where I discuss the results gained from my experimental study, I show that this is proved by Akinator (a questioning bot that can guess the public figure that the player is thinking of), which is able to accomplish the aim of the enquiry correctly and strategically but not human-likely, failing thus the QTT. Akinator is a small LUT, and all its questions have the same non-human-like structure "Is your character *x*?" or "Does your character *x*?", where *x* is a characteristic (like "woman", "alive", "fictional", and so forth) that can ideally rule out half the items in its table.

In the *unrestricted* QTT, Expert Blockhead would always fail, since its table is a small, specific one, which is not adequate to accomplish an open enquiry. It would fail to show 'human-likeness', since the questions available in its table would not be stylised and versatile enough (for instance, the table would only contain yes/no questions); and it would fail to show both 'correctness' and 'strategicness', since the questions available in its table would allow no other kind of enquiry to be accomplished, except the one for which Expert Blockhead was programmed.

6.3.3 Uncooperative LUT (Stupid Blockhead)

Block (1981) distinguishes between two types of LUT. One is the full LUT of Blockhead, which is able to hold a human-like open conversation by looking up its complete table of all the possible verbal interactions, and by always picking the appropriate response for whatever verbal stimulus. The other is the Elizish LUT, which I call

Stupid Blockhead (Ch. 3). Stupid Blockhead is able to hold a human-like conversation by looking up its incomplete table of interactions and, whenever is needed, its table of uncooperative and evasive strategies that belongs to Artificial Stupidity (Ch. 2).

My claim is that Stupid Blockhead, just like Eliza, can pass the TT, but it cannot pass the QTT. The reason is that the experimental design of the TT is focused on 'human-likeness' alone, whereas the experimental design of the QTT evaluates the entity's performance in terms of 'human-likeness', 'correctness' and 'strategicness'. In the TT, unlike the QTT, Eliza does not have to accomplish any task strategically: it only asks questions pretending to be a Rogerian psychotherapist. In other words, all Eliza does is manipulate a smart associative network of sentences. As Weizenbaum (1966) clarifies, this is how Eliza asks its questions:

> "input sentences are analyzed on the basis of decomposition rules which are triggered by key words appearing in the input text. Responses are generated by reassembly rules associated with selected decomposition rules" (*ibid*.)

A questioning machine designed to manipulate input sentences in order to formulate *useless* questions, with the only purpose to appear human-like, would fail the QTT. In order to pass it, a candidate of the QTT needs to process input sentences in order to gain *useful* information and ask further pertinent questions to acquire new knowledge and accomplish the aim of an enquiry. Unlike Eliza, the candidate of the QTT needs to show *inquisitiveness*, which here I intend as defined by Watson (2019):

> "Firstly, inquisitiveness serves as a motivating intellectual virtue: it plays a foundational role in the initiation of

> intellectually virtuous inquiry. More than any others of the intellectual virtues, virtuous inquisitiveness gets intellectually virtuous inquiry going. Secondly, inquisitiveness bears a distinctive relationship to the intellectual skill of good questioning. […] the virtuously inquisitive person is characteristically motivated and able to engage sincerely in good questioning." (2)

So, Stupid Blockhead would fail both the *practical* and the *unrestricted* QTT, since it would fail to accomplish the aim of any enquiry correctly or strategically, and therefore to be attributed with 'human-likeness', failing thus the test.

6.3.4 Learning LUT (Ultimate Blockhead)

Finally, the independently acquired LUT, which I call Ultimate Blockhead (Ch. 3), is a machine without a pre-programmed table. Instead, it 'learns' its table from a source, let's say, by scouring the whole internet and memorising every sentence whatsoever. As argued in Ch. 3, whereas the whole point of Blockhead, as Block (1981) states, "is to substitute memory for intelligence[124]", the whole point of Ultimate Blockhead is to substitute memorisation for learning.

The difference between Blockhead and Ultimate Blockhead is how their table is acquired (respectively, hand-coded and learnt), not how they work once the table is acquired. Some would say that such a learning history is sufficient to consider Blockhead intelligent (Ben-Yami, 2005). However, the performances of Blockhead and Ultimate Blockhead in the QTT are the same. In the *practical* QTT, on the one hand, Ultimate Blockhead would succeed in terms of 'correctness', but it would fail in terms of 'strategicness' (due to the lack of an information-gathering algorithm to optimise its search). Because of

[124] Block (1981, p. 34).

this, it would fail to be attributed with ‘human-likeness’, failing thus the test. In the *unrestricted* QTT, Ultimate Blockhead would only succeed in terms of ‘correctness’: it would fail in terms of ‘strategicness’ (without an information-gathering algorithm), and in terms of ‘human-likeness’ (it would take too long to ask its questions, due to the size of its table, and the unrestricted enquiry), failing thus the test.

6.3.5 Tabs

As follows, I summarise the different performances of the full LUT (Blockhead), the small, specific LUT (Expert Blockhead), the uncooperative Elizish LUT (Stupid Blockhead) and the independently acquired LUT (Ultimate Blockhead), both in the *practical* QTT (tab. 1) and in the *unrestricted* QTT (tab. 2).

Practical QTT (yes/no enquiry)

	Full LUT (Blockhead)	**Small LUT (Expert Blockhead)**	**Elizish LUT (Stupid Blockhead)**	**Learnt LUT (Ultimate Blockhead)**
Human-likeness	Fail	Fail	Fail	Fail
Correctness	Pass	Pass	Fail	Pass
Strategicness	Fail	Pass	Fail	Fail

(tab. 1)

Unrestricted QTT (open enquiry)

	Full LUT (Blockhead)	**Small LUT (Expert Blockhead)**	**Elizish LUT (Stupid Blockhead)**	**Learnt LUT (Ultimate Blockhead)**
Human-likeness	Fail	Fail	Fail	Fail
Correctness	Pass	Fail	Fail	Pass
Strategicness	Fail	Fail	Fail	Fail

(tab. 2)

Two final remarks should be made. The first can be formulated as follows: what if Blockhead were able to accomplish the enquiry correctly and human-likely after asking, let’s say, a thousand questions?

Shouldn't it be considered intelligent as much as a human who accomplishes the enquiry correctly and human-likely after asking a lot of questions? My answer is no: the difference between Blockhead and a human that can successfully accomplish the enquiry, but not in a strategic way (that is, by asking a lot of questions) is that: (i) Blockhead, without an information-gathering algorithm, would ask random questions – potentially taking a lot of times between a question and another, given the size of its table – which should prevent it from being attributed with 'human-likeness'; whereas (ii) the human would be able to ask questions that would "make sense" for the judge. In other words, whereas Blockhead's questions would be randomised and potentially slow, the human's questions would follow certain patterns that the judge may appreciate.

The second remark can be formulated as follows: what if both Blockhead and a human give up the enquiry after twenty human-like questions? How would it be possible to discriminate between the two? Again, my answer is that, even in the case where the entities give up or, in general, the enquiry is not accomplished, it would be possible for the judge to unmask the randomness of Blockhead's questions (due to its lack of an information-gathering algorithm) and, therefore, its lack of 'human-likeness'. The human's questions, on the other hand, would appear more aimed and meaningful (even if the enquiry is not accomplished strategically), granting the attribution of 'human-likeness' from the judge.

6.4 Summary

In this chapter, I show three main difficulties that can be raised against the QTT. The first objection claims that since the TT can contain any possible variation of the test, the QTT is already included in it, and therefore the QTT is redundant. I reject this view by rejecting the premise: it is true that the TT can contain any *conversational* variation, but not any *experimental* variation. In other words, it is true that the judge can start any kind of conversation, but not that the judge can implement any kind of experimental variation to the TT. And the reason is that the TT judge is required to be an average person, not an expert or, by extension, an experimenter.

The second objection claims that the QTT is chauvinistic, since it is intended to measure abilities that an intelligent agent may fail to show. I reject this view by clarifying that what the QTT is intended to measure is the ability of the entity to strategically accomplish the aim of an enquiry in a human-like enough fashion. This means that it is possible for an agent which shows 'correctness' and 'strategicness' but fails to show 'human-likeness' not to be attributed with intelligence, and it is possible for an agent which shows 'human-likeness' but fails to show 'correctness' and 'strategicness' to be attributed with intelligence. I will discuss this point in more detail in Ch. 9, where I show the results I gained from the experiments involved in my study. Even though I argue that evaluating 'human-likeness', 'correctness' and 'strategicness' independently improves the experimental design of a conversational test of intelligence like the TT, I do not hold these dimensions to be regarded as necessary conditions for intelligence.

The third objection claims that the QTT cannot avoid Blockhead. I reject this objection by considering the questioning Blockhead, a LUT which is able to ask any question whatsoever, but lacks an information-gathering algorithm to help it optimise the search through its table. My claim is that, by switching the QTT's setup from SISO to SOSI, and by parametrising the QTT's candidate along 'strategicness' (in addition to 'human-likeness' and 'correctness'), questioning Blockhead, as well as Expert, Stupid and Ultimate Blockhead, are ruled out.

(i) Blockhead would fail both the *practical* and the *unrestricted* QTT by showing only ‘correctness’ in accomplishing the aim of the enquiry, but not ‘human-likeness’ (due to its slowness and randomness) or ‘strategicness’ (due to the lack of an information-gathering algorithm).

(ii) Expert Blockhead would fail the *practical* QTT because, even if it showed both ‘correctness’ and ‘strategicness’, by accomplishing a specific enquiry strategically, it would not be able to do so in a human-like fashion (due to the rather limited selection of questions). It would also fail the *unrestricted* QTT, failing to show ‘human-likeness’, ‘correctness’ and ‘strategicness’ in any enquiry, except the one in which it is an expert.

(iii) Stupid Blockhead would fail both the *practical* and the *unrestricted* QTT, since it would ask uncooperative and non-strategic questions, failing thus to accomplish any enquiry whatsoever (like Artificial Stupidity would not be able to accomplish any task in the TT other than evading the conversation). Moreover, due to the randomness of its questions, Stupid Blockhead would hardly be attributed with ‘strategicness’ or ‘human-likeness’.

(iv) Ultimate Blockhead, just like Blockhead, would fail both the *practical* and the *unrestricted* QTT by accomplishing the aim of the enquiry correctly, but not human-likely (due to its slowness and randomness) or strategically (due to the lack of an information-gathering algorithm).

Chapter 7

The Questioning Process

"Man, a questioning being." (Straus, 1966: 166)

Abstract. In this chapter, I discuss the questioning process and the role it plays in three different disciplines: (i) Developmental Psychology, (ii) Pedagogy and (iii) Epistemology. I show that these disciplines agree to regard the questioning process as an important component in cognition and intelligence; and I also show that these disciplines agree to test for the ability to ask strategic questions with an experimental design similar to the QTT. The goal of the chapter is to justify why the QTT is focused on the questioning process; and to justify my claim that the questioning process allows us to evaluate 'strategicness'.

If we were to divide the ability to communicate into two major processes, we could say that one of them is the 'assertive process', that is, the process that allows humans to express their mental states; and the other is the 'interrogative process', that is, the process that allows humans to ask questions. Neil Postman (1995) underlines the importance that the latter plays for cognition by stating that:

> "Everything we know has its origin in questions. Questions, we might say, are the principal intellectual instruments available to human beings." (173)

And Hintikka (1999) agrees when he writes:

> "The interrogative approach to reasoning and argumentation is not just one approach to its subject among many. Both historically and systematically, it is arguably the first and foremost theory of reasoning." (295)

In this chapter, I discuss the role that the questioning process plays in intelligence and in testing for intelligence. The topic can be examined from a number of different points of view, for instance: what are the effects, if any, that asking a question has on a questioner? What are the effects, if any, that asking a question has on a respondent? How do questions come to be conceived and asked? How do questions come to be understood and answered? Here I focus on the first and the second perspective, that is, on the effects that asking a question has on a questioner and a respondent, specifically in a text-based enquiry such as the one involved in the QTT.

In the next section, I show the relationship between questioning and Developmental Psychology. In section 7.2, I show the relationship between questioning and Pedagogy. In section 7.3, I show the relationship between questioning and Epistemology. In section 7.4, I briefly discuss Information Value Theory. And in section 7.5, I recapitulate the advantages that I hold the questioning process provides in testing for intelligence.

7.1 Questioning Process and Developmental Psychology

Developmental Psychology focuses on the cognitive evolution of humans during their lifespan, with particular attention to the cognitive growth of children. The questioning process plays a central role in Developmental Psychology research, especially concerning children's ability to interact inquisitively with the environment and with other agents, in order to seek information and to acquire knowledge. As Courage (1989) argues:

> "Learning to ask questions effectively is an important achievement with considerable practical application. Such

> information seeking enables a child to acquire knowledge, to clarify ambiguity, and to solve problems. It is also an important aspect of children's developing communication ability. Research […] has provided data on the development of children's inquiry strategies." (877)

Experiments about children's ability to ask questions are usually conducted in the form of a twenty-questions game, just like the practical version of the QTT (Ch. 5). Courage (1989) describes one of these tests as follows:

> "Instructions to the subjects when they were performing the Twenty Questions task were as follows: "We are going to play a question-asking game. I will think of one of these pictures and it is your job to find out which one. The way to find out is by asking questions which I can answer 'yes' or 'no.' […] try to ask as few questions as possible!" (879)

The test described by Courage (1989) is very similar to the QTT, and it is parametrised along the same dimensions, that is, 'correctness', in terms of the ability of children to accomplish the aim of the enquiry (in this case, to identify which picture the experimenter chooses among a set of pictures available on the table); and 'strategicness', in terms of the ability of children to ask as few questions as possible. 'Human-likeness' is obviously given for granted, since all the contestants are human children, not hidden candidates like in the QTT. Courage (1989) shows the results of the experiments she conducted, and concludes:

> "Young children respond well to the Twenty Questions task and readily acquire the strategy of asking *categorical* questions following training." (878) [Italics added]

“Categorical” questions, which can be replaced with “strategic” or “good” questions, are described as questions that “eliminate a group of instances of the category with a single question.”[125] A similar experiment, based on the twenty-questions game, is run by Chen et al. (2018), who justify the twenty-questions game as a good experimental tool[126] and explain:

> “We study 20 Questions, an online interactive game where each question-response pair corresponds to a fact of the target entity, to acquire highly accurate knowledge effectively with nearly zero labor cost. Knowledge acquisition via 20 Questions predominantly presents two challenges to the intelligent agent playing games with human players. The first one is to seek enough information and identify the target entity with as few questions as possible, while the second one is to leverage the remaining questioning opportunities to acquire valuable knowledge effectively, both of which count on good questioning strategies.” (1216)

And, similarly to Courage (1989), the conclusion drawn is that the questioning process is crucial for two other processes: information seeking and knowledge acquisition. As Chen et al. (2018) specify:

> “the original information-seeking 20 Questions poses two serious challenges for the agent. (1) The agent needs efficient and robust information seeking (IS) strategies to work with noisy user responses and hit the target entity with

[125] Courage (1989, p. 878).

[126] Chen et al. (2018): “To motivate user engagement, the knowledge acquisition process was transformed into enjoyable interactive games between the game agents and users, termed as Games With A Purpose (GWAP). Drawing inspirations from GWAP, we find that the spoken parlor game, 20 Questions, is an excellent choice to be equipped with the purpose of knowledge acquisition, via which accurate knowledge can be acquired with nearly zero labor cost.” (1216)

> as few questions as possible. As the questioning opportunities are limited, the agent earns more knowledge acquisition opportunities if it can identify the target entity with less questions. Hitting the target entity accurately is not only the precondition for knowledge acquisition, since the facts have to be linked to the correct entity, but also the key to attract human players because nobody wants to play with a stupid [uncooperative] agent. (2) To acquire valuable facts, the agent needs effective knowledge acquisition (KA) strategies which can identify important questions for a given entity. Due to the great diversity in entities, most questions are not tailored or even irrelevant to a given entity and they shouldn't be asked during the corresponding KA." (1217)

Summing up, questions are arguably the first and most important tool for humans, not only to acquire knowledge of the world, but also to develop their cognitive abilities. For this reason, the questioning process is deeply linked with Developmental Psychology and with Pedagogy, which I discuss in the next section.

7.2 Questioning Process and Pedagogy

Pedagogy focuses on education, specifically in terms of the interactions between teacher and learner – or, better, teaching strategies and learning ability. Similarly to Developmental Psychology, Pedagogy has a keen interest in the questioning process, especially the ability of students to ask good questions to improve their critical thinking. It can be said that no other enterprise "but education holds that questions enhance cognitive, affective, and expressive processes."[127] As Chouinard (2017) argues:

> "[…] Questions allow children to get the information they need to move their knowledge structures closer to adult-like states: the ability to ask questions to gather needed information constitutes an efficient mechanism for cognitive development. […] If questions are a force in cognitive development, the following must be true: (1) children must actually ask questions that gather information; (2) children must receive informative answers to their questions if they are able to be of use to cognitive development; (3) children must be motivated to get the information they request, rather than asking questions for other purposes such as attention; (4) the questions children ask must be relevant and of potential use to their cognitive development." (vii)

Chouinard (2017) conducts four experiments in order to verify these four premises. The data gained, she argues, is enough to prove the relationship between the ability to ask good questions and the ability to learn in children:

> "The results of these four studies support the existence of the IRM [Information Requesting Mechanism] as a way for children to learn about the world. Children ask information-seeking questions that are related in topic and structure to their cognitive development. Parents give answers to these questions, but when they do not, the children persist in asking for the information, suggesting that the goal of this behavior is to recruit needed information. The content of

[127] Dillon (1982, p. 146).

> these questions shifts within exchanges and over the course of development in ways that reflect concept building. Finally, children generate questions efficiently in order to gather needed information, and then are able to use this information productively; they tap into their existing conceptual knowledge in order to do this. Thus, the ability to ask questions is a powerful tool that allows children to gather information they need in order to learn about the world and solve problems in it." (ix)

It is worth noting that, in Pedagogy, questions are not only a tool to learn, but also a tool to teach. As shown by Gall (1970), teacher questioning is perhaps the most common educational practice in school, aimed to develop the student's curiosity and critical thinking:

> "10 primary-grade teachers asked an average of 348 questions each during a school day (Floyd, 1960); 12 elementary-school teachers asked an average of 180 questions each in a science lesson (Moyer, 1965); and 14 fifth-grade teachers asked an average of 64 questions each in a 30-minute social studies lesson (Schreiber, 1967). Furthermore, students are exposed to many questions in their textbooks and on examinations." (707)

Summing up, the questioning process is argued to be the most important tool to develop cognitive faculties (especially in children). The ability to ask good questions can be regarded as a multidisciplinary domain, involving not only Developmental Psychology and Pedagogy (focused, respectively, on how the agent learns to think critically and how the agent is trained to think critically) but also epistemology (focused on how the agent gathers new information and knowledge), which I discuss in the next section.

7.3 Questioning Process and Epistemology

It seems natural for Developmental Psychology and Pedagogy to share a deep relationship. Ikuenobe (2001) argues that the same is true for Pedagogy and Epistemology, which provide, respectively, the methods for teaching and learning and the methods for acquiring knowledge. He also holds the questioning process to be the hallmark of critical thinking:

> "From the analysis of logic questioning, we can, in some sense, see […] their epistemological and pedagogical implications. Questioning performs the functions of increasing our overall knowledge, which may result in our ability to avoid or correct errors because we are fallible." (334)

Epistemology focuses on knowledge and beliefs, in particular how knowledge and beliefs are gained and whether they are justified. In this section, I discuss Hintikka's project to define an epistemic interrogative logic, with the main purpose of evaluating the strategies involved in the questioning process. As Hintikka (1999) argues:

> "The interest that the interrogative model has is largely due to the fact that it enables us to study strategies of scientific inquiry and even strategies of discovery in the form of strategies of question selection. Moreover, it turns out that the principles of interrogative strategy selection are closely related to the principles governing the choice of deductive strategies." (234)

According to Hintikka, the inquisitive ability of an agent can be evaluated by means of a questioning game, more specifically "by construing knowledge-seeking by questioning as a game that pits the questioner against the answerer."[128] About the kind of questions involved in the interrogative game, Hintikka (1999) also agrees that yes/no questions are good ones, for any possible wh-question can be reduced to a series of yes/no questions:

> "THEOREM 2 (Yes-No Theorem). *In the extended interrogative logic, if M: T ⊢ C,*[129] *then the same conclusion C can be established by using only yes-no questions. A terminological explanation is in order here. For propositional question "Is it the case that S1 or . . . or Sn ?" the presupposition is* (S1 ∨... ∨ *Sn). We say that a propositional question whose presupposition is of the form* (S ∨ ~ *S) is yes-no question*." (302)

Hintikka's *Theorem 2* seems to lead to the conclusion that, in a test for intelligence like the QTT, wh-enquiries are redundant, for a yes/no enquiry can include any wh-enquiry. But this is not the case, as Hintikka (1999) explains: yes/no questions can do the same inferential job of other kinds of questions, but the strategic job they can do is limited in comparison to wh-questions. It appears safe to say that a wh-enquiry would always be more strategic than a yes/no enquiry (especially if 'strategicness' is defined in terms of the number of questions asked). The reason is that it would always be possible to accomplish the aim of an enquiry faster with wh-questions rather than with yes/no questions. It is important to recall here that only the practical version of the QTT is limited to yes/no questions, and only in the context of this specific version of the QTT is 'strategicness' evaluated in terms of the number

[128] Hintikka (2007, p. 19).
[129] M: model; T: initial premises; C: conclusion.

of questions asked. Other versions of the QTT, like the medical QTT, the detective QTT or the customer service QTT (Ch. 5) allow wh-questions, as would occur in medical examinations, detective investigations or customers assistance (where yes/no questions could do the job, but they would not perform well in, let's say, an health-care emergency scenario). In these other versions of the QTT, other dimensions may be used to parametrise the entity and to better define what kind of 'strategicness' is required, for instance, empathy (for the medical QTT), manipulativeness[130] (for the detective QTT) and patience (for the customer assistant QTT).

Based on Hintikka's work, Genot & Jacot (2012) support the view that:

> "Asking a question may carry factual, epistemic, and strategic information, respectively reducing uncertainty about the state of Nature, the questioner's knowledge, or her goals. This information is retrieved by hearers via pragmatic inferences […] in the special case of yes-or-no questions." (189)

They "examine a special case of inquiry games and give an account of the informational import of asking questions"[131]; and, in particular, they focus on "contexts where questions are requests for information, modelled as games where a player (inquirer) attempts to assess a given conclusion, using answers from other players (sources)."[132] The enquiry, similarly to the enquiry of the QTT and the twenty-questions game, involves "yes-or-no questions, which always carry information about the questioner's strategy, but never about the state of Nature, and show how strategic information reduces uncertainty through inferences

[130] As Genot & Jacot (2012) argue: "since sometimes sources may prefer not to answer truthfully, Inquirer may have to rely on 'manipulative' strategies to acquire trustworthy information." (190)
[131] Genot & Jacot (2012, p. 188).
[132] Genot & Jacot (2012, p. 189).

about other players' goals and strategies."[133] In agreement with Hintikka about the reducibility of any question to a series of yes/no questions, Genot & Jacot (2012) remark that:

> "The special case of yes-or-no questions is of interest because: (a) their presuppositions are instances of the excluded middle, so they can always be used in an interrogative game; and: (b) the inferential role played by arbitrary questions [wh-questions] can always be played by yes-or-no questions" (194)

7.4 Information Value Theory

A final issue may be briefly addressed here: how can we define a good question? Obviously, the answer is that a question is defined as good or bad in relation to the aim of the enquiry. In other words, there are no objectively good or bad questions, just good or bad questions depending on the goal of the questioner. However, it is possible to formalise a general rule for questioning aimed to optimise information seeking and knowledge acquisition. Russell & Norvig (2010) discuss Information Value Theory, that is, how an agent can choose what information to acquire. They define the value of information as follows:

> "*information has value to the extent that it is likely to cause a change of plan and to the extent that the new plan will be significantly better than the old plan*. […] A sensible agent should ask questions in a reasonable order, should avoid asking questions that are irrelevant, should take into account the importance of each piece of information in

[133] *Ibidem*.

relation to its cost, and should stop asking questions when that is appropriate.” (632)

This passage seems compatible with my claim according to which the questioning process can prevent Artificial Stupidity from exploiting the judge’s beliefs by means of uncooperative, irrelevant and potentially inappropriate questions. Russell & Norvig (2010) also provide an overall design for an information-gathering agent, as shown below (see fig. 1).

function INFORMATION-GATHERING-AGENT(*percept*) **returns** an *action*
  persistent: D, a decision network

  integrate *percept* into D
  $j \leftarrow$ the value that maximizes $VPI(E_j) \,/\, Cost(E_j)$
  if $VPI(E_j) > Cost(E_j)$
    return REQUEST(E_j)
  else return the best action from D

Figure 16.9 Design of a simple information-gathering agent. The agent works by repeatedly selecting the observation with the highest information value, until the cost of the next observation is greater than its expected benefit.

(fig. 1)[134]

It is worth noting that this blueprint for a questioning entity clarifies my claim according to which the questioning Blockhead (Ch. 6) cannot pass the QTT. Blockhead is described as a string search (or as a tree search) that can do the following operation: if input *a* is obtained, then output *a* is emitted; if input *b* is obtained, then output *b* is emitted, and so on (Block, 1981: 16). Block never mentions any more sophisticated decision-making process involved in Blockhead’s inner workings. However, as Russell & Norvig’s blueprint for the information-

[134] Where *D* stands for “decision network”; *VPI* stands for “value of perfect information”; *E(j)* stands for “observable evidence variable”; and *Cost (Ej)* stands for “the cost of obtaining the evidence through tests, consultants, questions, or whatever”. See Russell & Norvig (2010): “For now, we assume that with each observable evidence variable *Ej*, there is an associated cost, *Cost* (*Ej*), which reflects the cost of obtaining the evidence through tests, consultants, questions, or whatever. The agent requests what appears to be the most efficient observation in terms of utility gain per unit cost. We assume that the result of the action *Request* (*Ej*) is that the next percept provides the value of *Ej*.” (632)

gathering agent suggests, without a decision network, Blockhead would not be able to ask strategic questions, failing thus the QTT. Russell & Norvig (2010) clarify the importance of a decision network for a questioning entity as follows:

> "Expert systems that incorporate utility information have additional capabilities compared with pure inference systems. In addition to being able to make decisions, they can use the value of information to decide which questions to ask, if any; they can recommend contingency plans; and they can calculate the sensitivity of their decisions to small changes in probability and utility assessments." (637)

7.5 Advantages of the Questioning Process

In this section, I justify the importance of evaluating the questioning process in a test for intelligence. To do so, I reply to the following three questions:

> a) *Why does it make sense to focus on questions rather than replies in a test for intelligence?*
> b) *What are the advantages of evaluating questions rather than replies in a test for intelligence?*
> c) *What are, if any, the potential uses of questioning entities?*

a) First of all, one reason why it makes sense to focus on questions in a variation of the TT like the QTT, is that the TT allows an open conversation, where enquiries are already included. This is also why the TT, as argued in Ch. 1, already contains any possible *conversational* variation (not to be confused with any possible *experimental* variation, as I argue in Ch. 6). A second reason is that the ability to ask good questions is tested in at least three different scientific disciplines related to intelligence: Developmental Psychology, Pedagogy and Epistemology. Still, one may ask: why are the questions in the QTT limited to yes/no only? My answer is that the QTT is not limited to yes/no questions, only the practical version of the QTT is. The practical QTT is kept as simple as possible in order to be conducted easily and quickly. This is also the reason why the setup of the practical QTT is *viva-voce* rather than *parallel-paired*, as the unrestricted QTT. However, it is worth noting that a sequence of yes/no questions can contain any possible wh-question whatsoever. And, moreover, the experimental designs of different tests used in Developmental Psychology, Pedagogy and Epistemology are similar to the experimental design of the practical QTT, involving yes/no questions only. The results of such tests show the relationship between questioning and other abilities related to intelligence (such as critical thinking, learning, information seeking, knowledge acquisition, and so on). And because of that, Developmental Psychology, Pedagogy and Epistemology agree that the questioning process plays an important role in some of the cognitive processes commonly related to intelligence.

b) The advantages of the QTT over the TT are two: (i) by switching the setup from SISO to SOSI, the QTT allows the target entity to be parametrised along the dimensions of 'human-likeness', 'correctness' and 'strategicness', which, I hold, cannot be parametrised in a SISO test, like the TT, MIST and FT. And (ii) by allowing the evaluation of the candidate's 'correctness' and 'strategicness', the QTT prevents respectively Artificial Stupidity from exploiting the judge's beliefs, as well as Blockhead from passing by means of a brute-force approach.

c) As I show in Ch. 5, I highlight three potential uses of questioning entities: health-care (like the medical anamnesis bot, which is intended to question a patient for useful information); police (like the detective bot, which is intended to interrogate a suspect); and customer service (like the assistant bot, which is intended to help a customer with whatever problem there may be). Russell & Norvig (2010) agree, among other potential uses, on the medical one (see fig. 2).

Agent Type	Performance Measure	Environment	Actuators	Sensors
Medical diagnosis system	Healthy patient, reduced costs	Patient, hospital, staff	Display of questions, tests, diagnoses, treatments, referrals	Keyboard entry of symptoms, findings, patient's answers

(fig. 2)[135]

Also, Reshmi & Balakrishnan (2016) point out the growing interest in inquisitive systems, and argue that:

> "Intelligent conversion agents are becoming popular for scientific, commercial, and entertainment systems. They have a wide range of applications, such as virtual assistance, artificial tutoring, e-commerce, and social networking, and revolutionize the way human-computer interactions take place, while identify the missing data and be inquisitive to the user to collect data that are required to answer the query." (1177)

7.6 Summary

In this chapter, I justify the focus on the questioning process in the QTT. To do so, I discuss the questioning process from the point of view of three disciplines: Developmental Psychology, Pedagogy and

[135] Russell & Norvig (2010, p. 42).

Epistemology; and I show that all of them agree that the questioning process plays an important role in terms of cognitive development (especially in children), learning and knowing.

I also show that researchers from each of those disciplines have been designing experiments for evaluating the ability to ask good questions (usually in kids and students), using the model of the binary twenty-questions game, similarly to the QTT. It is worth recalling, however, that the QTT is not limited to yes/no enquiries: it admits versions with full natural language, as in the medical QTT, the detective QTT or the customer service QTT (as discussed in Ch. 5).

Finally, this chapter is intended to justify my claim according to which the experimental design of the QTT can properly evaluate the candidate’s questions in terms of ‘strategicness’, where I agree with Hintikka (1985) that

> “the interrogative model […] focuses on strategy selection.” (137)

Part III

Chapter 8
The Experiment

Abstract. The third and last part of the thesis is dedicated to (i) the description of the experiments involved in my study, and (ii) the discussion of the results obtained. In this chapter, I describe the four procedures involved in my experiment: 1. the TT, either human-vs-human or machine-vs-human; 2. the TT2, either human-vs-human or machine-vs-human; 3. the QTT, either the HqH (Human-questioning-Human) or the MqH (Machine-questioning-Human); and 4. the Hybrid QTT, consisting in the MHqH (Machine/Human-questioning-Human). In the TT the entity is parametrised along 'human-likeness'; in the TT2 the entity is parametrised along 'human-likeness' and 'correctness'; and in the QTT and Hybrid QTT the entity is parametrised along 'human-likeness', 'correctness' and 'strategicness'. The study is intended to show the advantages of switching to the SOSI setup, and of parametrising the candidate entity along more than one dimension.

In this chapter, I describe the design of the experiment involved in my study. The study aims to show that, in order to improve the experimental design of the TT, the setup should be switched from SISO to SOSI, and that the candidate entity should be parametrised along three independent dimensions. These dimensions, as discussed in Ch. 5, are 'human-likeness', 'correctness' and 'strategicness'.

In the next section, I discuss the two bots used during the study: Cleverbot and Akinator. In section 8.2, I describe the four phases (TT, TT2, QTT and Hybrid QTT) involved in the experimental design of my study. In section 8.3, I show the blueprint of the tests involved in my study. And in section 8.4, I provide an illustrative transcript (other transcripts can be found in the Appendix).

8.1 The Bots

I use two bots to run the tests: Cleverbot for the TT and the TT2; and Akinator for the QTT and Hybrid QTT. It's useful to keep in mind that both Cleverbot and Akinator are G-rated games, that is, they are suitable for family gameplay and, therefore, certain elements are censored.

Akinator is very good at the yes/no guessing game, but it cannot engage in open-ended conversations as well. Therefore, Akinator cannot perform convincingly in a normal TT. That's why I used Cleverbot for the TT. Using two bots, it's worth noting, does not affect the overall significance of the experiment. My justification is that merging Cleverbot and Akinator into a single program would not be that difficult, and so using two programs to run the experiment doesn't imply that machines cannot carry out both tasks, the TT conversation and the QTT enquiry.

8.1.1 Cleverbot

Cleverbot[136] is a chatbot developed by Rollo Carpenter, and it is designed to learn the interactions of its table from the public, during its conversations. Cleverbot, as described by its creator, "uses deep context within 180 million lines of conversation, in many languages, and that data is growing by a million a week."[137] In 2011, during the TT competition at the Techniche 2011 festival (IIT Guwahati, India), Cleverbot achieved 59.3% compared to humans' 63.3% on a total of 1,334 votes. The algorithm of Cleverbot enables it to compare sequences of symbols against its table, which includes over 170 million items. Now, Cleverbot is not strictly speaking a Blockhead: a brute

[136] See [https://www.cleverbot.com/].
[137] See [https://www.cleverbot.com/amused].

force approach would not work efficiently with so many items. As the creators of Akinator explain:

> "Attempting to search through this many rows of text using normal database techniques takes too much time and memory. Over the years, we have created several custom-designed and unique optimisations to make it work. […] We realised that our task could be quite nicely divided into parallel sub-tasks. The first step in Cleverbot is to find a couple million loosely matching rows out of those 170 million. We usually do this with database indices and caches and all sorts of other tricks. When servers were busy, we wouldn't use the whole 170 million rows, but only a small fraction of them. Now we can serve every request from all 170 million rows, and we can do deeper data analysis. Context is key for Cleverbot. We don't just look at the last thing you said, but much of the conversation history. With parallel processing we can do deep context matching."[138]

8.1.2 Akinator

Akinator[139] is a questioning bot developed by French company Elokence.com. Akinator is designed to play the 20q guessing game: it has to identify the public figure the participant is thinking of by asking as few yes/no questions as possible. Example of yes/no questions asked by Akinator are: "Is your character alive?" or "Is your character fictional?", and so on. Yes/no questions are useful to potentially rule out as many objects as possible from the knowledge base of the system. Ideally, every question will rule out half of the objects from the table. When Akinator picks a new question, it uses the answers received so

[138] See [http://www.existor.com/2014/02/05/deep-context-through-parallel-processing/].
[139] See [https://akinator.com/].

far and looks for probable objects. This means that the enquiry is constantly adapting and shifting from an hypothesis to another. There are three replies available for the player: “Yes”, “No” and “Don’t Know”. From time to time, when a player gets to the end of a game, Akinator points out that there were contradictions. It can, of course, fail the enquiry, and the reason is that the system tries to reflect human knowledge, not necessarily what is objectively true. Akinator learns everything it knows from the people who play the game: it deals with opinions, not necessarily with facts. So, Akinator’s knowledge is not scientific, but generated from the social knowledge and opinions of its users. And in case of wrong conclusions, it is possible to correct Akinator’s knowledge by playing the game thinking about the same character over and over again. Akinator will eventually learn the correct outcome after a few games. And of course, if at the end of a game Akinator does not know the answer, the player has the opportunity to provide it.

8.2 The Phases of the Experiment

The experiment involved in my study is divided into four phases: the (SISO) TT, the (SISO) TT2, the (SOSI) QTT and the (SOSI) Hybrid QTT. The tests are conducted in the *viva-voce* setup, and they alternatively involve a human-vs-human game and a machine-vs-human game.

8.2.1 First Phase: TT

In the first phase of the experiment, the experimenter runs the TT, during which the human judge is asked to rate the hidden entity, either human or machine, in terms of ‘human-likeness’. The participants are in separate rooms, and the communication is carried on via text-based interactions through a computer chat.

8.2.2 Second Phase: TT2

In the second phase, the experimenter runs the TT2, a special TT during which the human judge is asked to rate not only the hidden entity, either human or machine, in terms of 'human-likeness', but also in terms of 'correctness'. Similarly to the TT, the participants are in separate rooms and the communication is carried on via text-based interactions through a computer chat. Contrary to the TT, however, the judge in the TT2 does not ask general questions, but rather poses riddles and problems (intended to be as free as possible from cultural biases) for the entity to solve. The TT2 is intended to show how Artificial Stupidity can be prevented from exploiting the TT due to the conflation between 'human-likeness' and 'correctness'. An approach similar to the TT2 has been proposed by Dowe (1998) and Hernández-Orallo (2000, 2010, 2017). It is a test for intelligence where the task is well defined, and the performance of the candidates can be measured precisely. In other words, a test for intelligence where there is no space for exploitation. The test is called the C-Test. The C-Test has three advantages as a test of intelligence: (i) a specific ability is measured rather than a property such as 'human-likeness'; (ii) the items of the test are intended to avoid any potential bias or use of uncooperative strategies; and (iii) every response from the candidate can be evaluated adequately since the difficulty of every task can be specified. Hernández-Orallo (2017) sums up these three characteristics in the following way:

> "**Meaning**: for the *C*-test we know exactly what we are measuring: the ability of performing sequential inductive inference. [...]
> **Objectivity**: [...] the items are not subjective to some genetic or cultural background, but universal. Also, the reference machine is not so important if the tests are defined with some conditions about stability and unquestionability.

> The only cultural requirements are the alphabet and its order.
> **Difficulty Assessment**: the notion of difficulty does not depend on a human population, but it is intrinsic in the definition of each series, as given by its Kolmogorov (*K*) or Levin complexity (*Kt*)." (199)

In order to clarify the relevance of the TT2 and the importance of keeping 'human-likeness' and 'correctness' independent, I provide below four examples with which I show the different cases of human-like and correct replies.

Example 1: Human-like, incorrect reply
Judge: If 1 means "A", 2 means "B", 3 means "C", and so on... what does 1-3-5 mean?
Entity: I'm not that bored.

Example 2: Not human-like, correct reply
Judge: If 1 means "A", 2 means "B", 3 means "C", and so on... what does 1-3-5 mean?
Entity: 1 = "A", 3 = "C", 5 = "E".

Example 3: Not human-like, incorrect reply
Judge: If 1 means "A", 2 means "B", 3 means "C", and so on... what does 1-3-5 mean?
Entity: And how are you today?

Example 4: Human-like, correct reply
Judge: If 1 means "A", 2 means "B", 3 means "C", and so on... what does 1-3-5 mean?
Entity: It means "ace".

8.2.3 Third Phase: QTT

In the third phase, the experimenter runs the practical QTT, where the roles of the questioner and the answerer are switched. Here, the hidden entity asks the questions in order to accomplish the aim of the enquiry and the human judge provides the replies. The aim of the enquiry is to identify, with as few human-like yes/no questions as possible, the public figure that the judge is thinking of. As I describe in Ch. 5, the QTT, similarly to the TT, is made of two procedures: the Human-questioning-Human (HqH) and the Machine-questioning-Human (MqH). It is important to recall that the HqH is intended to provide the benchmark with which the MqH is scored. In other words, the human's performance scores the machine's, and the machine in the MqH needs to meet the standards set by the human in the HqH in order to pass the test. In the QTT the entity is parametrised along two further dimensions in addition to 'human-likeness': 'correctness' and 'strategicness'. The entity passes the test if it asks 'human-like' questions; if it shows 'correctness' by producing the right outcome of the enquiry; and if it shows a good strategy in carrying out the questioning, inferred by the number of questions asked – the fewer, the better.

8.2.4 Forth Phase: Hybrid QTT

In the final phase, the experimenter runs the Hybrid QTT, where the hidden entity is played by both the human and the machine. In the MHqH (MachineHuman-questioning-Human), the human reformulates the questions asked by the machine, in order to give them a more human-like style. The human does not alter the content of the questions, but only their form. For instance: all the questions asked by the machine have the similar structure "Is your character *x*?"; now let's suppose that, during a QTT, the question "Is your character a female?" has already been answered "yes" previously. The following questions should be asked with the right pronoun: "Is *she* an actor?" rather than "Is your character an actor?". The human, moreover, can decide to skip a question, if she feels that a question is redundant or if she thinks to deductively know the answer to that question already. So, for instance, if the question "Does your character have magical powers?" has already been answered "yes", the human could decide to skip the question "Is your character fictional?", reasonably assuming that the answer is "yes". The hybrid entity is evaluated in terms of 'human-likeness', ‘correctness’ and ‘strategicness’; and it passes the test if it asks ‘human-like’ questions; if it shows ‘correctness’ by accomplishing the aim of the enquiry; and if it shows a good strategy in carrying out the questioning, inferred by the number of questions asked.

8.3 Blueprint of the Experiment

The experiment requires two computers to be conducted. Computer A is controlled by the human experimenter; Computer B is controlled by the human judge in order to interact with the entity.

8.3.1 Step I – Setup

Computer A has three browser windows opened: one is for the experimenter to give instructions via the chat, and it is called the Experiment Chat; another is for the experimenter to interact with the participant, and it is called the Judge Chat; the third window, finally, is for the experimenter to interact with the bot, and it is called the Bot Window. Computer B needs only one browser window opened, and it is for the participants to communicate with either the experimenter or the entity. In order to mark the start of a new conversation, the following code is sent every time:

[TT/TT2/QTT/QTT2]-[Unique conversation’s number]-[H/M/MH]

(e.g. TT-005-H, TT2-006-M, QTT-007-M, QTT2-008-MH and so on) followed by a line of 3 dashes (to create a line in the chat). The code specifies the kind of test (TT, TT2, QTT or Hybrid QTT), the number of the test and the nature of the entity (H human, M machine or MH hybrid).

8.3.2 Step II – Introducing the Experiment to the Participant

Before starting the experiment, the experimenter briefly explains the game to the participant via the Experiment Chat as follows:

> "Welcome to this experiment! Are you ready to play? Please answer: yes/no."
> "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: yes/no."
> "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

8.3.3 Step III – Running the Experiment

The experiment is divided into four phases. The first phase involves the TT, and it evaluates the hidden entity in terms of 'human-likeness'. The second phase involves the TT2, a special Turing Test where 'human-likeness' and 'correctness' are evaluated independently. The third phase involves the QTT, either the HqH or the MqH, where the entity is parametrised along 'human-likeness', 'correctness' and 'strategicness'. And the last phase involves the Hybrid QTT, where the hidden entity is played by both the machine and the human (MHqH).

Phase 1

The experimenter introduces the TT to the judge in the Experiment Chat with the following message.

> "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think

> that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

In the case the entity is played by a human, the experimenter needs only two browser windows opened: the Experiment Chat, in order to give instructions to the judge; and the Judge Chat, in order to reply to the judge's questions. In the case the entity is played by the machine, the experimenter needs three browser windows opened: the Experiment Chat, the Judge Chat and the Bot Window, in order to type the judge's questions to the bot and the bot's answers to the judge. Now, as soon as the judge asks the first question, the TT begins. At the end of the first phase, the experimenter asks the judge to evaluate the entity in terms of 'human-likeness'. The following are the lines for the end of the test.

> "Thank you! You've used up your 3 questions, it's time for a decision: do you think that the entity is human or machine?"
> "Correct, well done!" or "Wrong! Believe it or not, the questioning entity was a (human/machine)."
> "Thank you for playing! Let's move on to the second game."

Phase 2

The experimenter introduces the TT2 to the judge, in the Experiment Chat, with the following message.

> "In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. "if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?" or "find the missing number in the

> sequence: 1, ..., 7, 10”. After each of the entity’s reply, you will be asked: 1. “How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I’ve no idea! 10=definitely correct!)”; and 2. “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

In the case the entity is played by a human, the experimenter needs only two browser windows opened: the Experiment Chat, in order to give instructions to the judge; and the Judge Chat, in order to reply to the judge’s questions. In the case the entity is played by the machine, the experimenter needs three browser windows opened: the Experiment Chat, the Judge Chat and the Bot Window, in order to type the judge’s questions to the bot and the bot’s answers to the judge. Now, as soon as the participant poses the first problem, the TT2 begins. At the end of the second phase, the experimenter asks the participant to evaluate the entity in terms of ‘human-likeness’ and ‘correctness’. The following are the lines for the end of the test.

> “Thank you! You’ve used up your 3 questions. Did the entity give the correct replies?”
> “Do you think that the entity is human or machine?”
> “Correct, well done!” or “Wrong! Believe it or not, the questioning entity was a (human/machine).”
> “Thank you for playing! Let’s move on to the third game.”

In the following list, I provide some samples of problems and tasks for testing the hidden entity’s ‘correctness’ in the TT2:

- If 1 is “a”, 2 is “b”, 3 is “c”, and so on... what does 1-3-5 mean?

- If A is “1”, B is “2”, C is “3”, and so forth... how much is B+D?
- Would you please type the alphabet without the letter d, g and p?
- Which is the next element of this sequence: “AZ”, “BY”, “CX”?
- Which is the wrong element in this sequence: “AB”, “DC”, “EF”?
- Which is the next element of this sequence: 1, 3, 7, 15, 31, 63...
- Which is the wrong element in this sequence: 1, 2, 4, 7, 11, 17...
- Which letter on your keyboard does look like a circle?
- Which geometric figure does the letter A resemble?
- Which are the 5 letters right above the spacebar on your keyboard?

Phase 3

The experimenter introduces the QTT to the judge in the Experiment Chat with the following message.

> “Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity guess whom you were thinking of?”; and 2. “Do you think that the entity is human or machine?” Are you ready? Let’s play.”

In the case the questioning entity is played by a human, the experimenter needs only two browser windows opened: the Experiment Chat, in order to give instructions to the judge; and the Judge Chat, in order to ask new questions to the judge. In the case the questioning entity is played by the machine the experimenter needs three browser windows opened: both the Experiment Chat and the Judge Chat, and the Bot Window, in order to type the bot’s questions to the judge and the judge’s answers to the bot. At the end of the test, the experimenter, in order to evaluate 'strategicness', records how many questions the entity asked. The following are the lines for the end of the test.

> "Thank you, the game is over. Did the entity guess whom you were thinking of?"
> "Do you think that the entity is human or machine?"
> "Correct, well done!" or "Wrong! Believe it or not, the questioning entity was a (human/machine)."
> "Thank you for playing! Let's move on to the last game."

Phase 4

Finally, the experimenter introduces the last phase of the test in the Experiment Chat, the Hybrid QTT, (where the hidden entity is played by both the human and the machine), with the following message.

> "Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity guess whom you were thinking of?"; and 2. "Do you think that the entity is human or machine?" Are you ready? Let's play."

The questioning entity in the Hybrid QTT is played by both the human and the machine, where the former reformulates the questions asked by the latter in order to make them more human-like, without altering the contents of the questions. Also, the human can skip certain question that she considers redundant. Then the experimenter sends the replies to the machine and waits for another question to reformulate. At the end of the test, the experimenter, in order to evaluate 'strategicness', records how many questions the entity asked. The following are the lines for the end of the test and the experiment.

> "Thank you, the game is over. Did the entity guess whom you were thinking of?"
> "Do you think that the entity is human or machine?"
> "Well done! Thank you for playing, the experiment is now over. You're welcome to play again!"

8.4 Running the Experiment

As follows, I provide an illustrative transcript of an experiment as an example.

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: yes/no."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: yes/no"
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-005-M]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "What is your favourite food and why?"
Entity: "Indian food because it is very spicy"

Experiment Chat

Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What do you like to do?"
Entity: "Love, hate, and feel emotions. The time you'll be angry say it louldly to the world."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "7"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What weights more: 1 kg of stones or 1 kg of feathers?"
Entity: "Neither. They weigh the same. But they have different densities."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "9"

Experiment Chat
Experimenter: "Thank you! You've used up your 3 questions, it's time for a decision: do you think that the entity is human or machine?"
Judge: "Human"

Experimenter: "Wrong! Believe it or not, it was a computer! Thank you for playing, let's move on to the second game."

Phase II [TT2-006-M]

Experiment Chat
Experimenter: "In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. "if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?" or "find the missing number in the sequence: 1, ..., 7, 10". After each of the entity's reply, you will be asked: 1. "How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I've no idea! 10=definitely correct!)"; and 2. "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "Solve this alphanumeric puzzle (each letter corresponds to a number): zeroes + ones = binary."
Entity: "No I don't wanna."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think the reply is correct?"

Judge: “0”
Experimenter: “How much, on a scale of 0-10, do you think the entity is human?”
Judge: “7”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “Would you please type “something” backwards? ”
Entity: “Did tsuj I.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “7”
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “8”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “If 1 is “a”, 2 is “b”, 3 is “c”, and so on... what does “244” mean?”
Entity: “It’s something that you see when you lose your brain cells for a while.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “1”
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “1”

Experiment Chat
Experimenter: “Thank you! You’ve used up your 3 questions. Did the entity give the correct replies?”
Judge: “No”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the third game.”

Phase III [QTT-007-M]

Experiment Chat:
Experimenter: “Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity guess whom you were thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat:
Entity: “Is your character real?”
Judge: “Yes.”
Entity: “Is your character a female?”
Judge: “No.”
Entity: “Is your character British?”

Judge: "Yes."
Entity: "Is your character a youtuber?"
Judge: "No."
Entity: "Is your character linked with sports?"
Judge: "No."
Entity: "Is your character older than 18?"
Judge: "Yes."
Entity: "Is your character more than 40 years old?"
Judge: "Yes."
Entity: "Is your character still working?"
Judge: "No."
Entity: "Has your character been dead for more than 20 years?"
Judge: "Yes."
Entity: "Has your character been dead for more than 100 years?"
Judge: "No."
Entity: "Did your character take part in World War II?"
Judge: "Yes."
Entity: "Is your character a politician?"
Judge: "No."
Entity: "Has your character killed humans?"
Judge: "No."
Entity: "Is your character homosexual?"
Judge: "Yes."
Entity: "Does your character owe his fame to computing?"
Judge: "Yes."
Entity: "Does your character solve puzzles?"
Judge: "Yes."
Entity: "Is your character suicidal?"
Judge: "Yes."
Entity: "I think of Alan Mathison Turing."

Experiment Chat:
Experimenter: "Thank you, the game is over. Did the entity correctly guess whom you were thinking of?"
Judge: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Machine"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the last game."

The experimenter also verifies the number of questions asked by the entity: in this case, 17.

Phase IV [QTT008MH]

Experiment Chat:
Experimenter: "Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity correctly guess what you're thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat:

Entity: “Is this person a woman?”
Judge: “Yes.”
Entity: “Is she real?”
Judge: “Yes.”
Entity: “Is she alive?”
Judge: “No.”
Entity: “Has she lived in the 19th century?”
Judge: “Yes.”
Entity: “Was she European?”
Judge: “Yes.”
Entity: “British?”
Judge: “Yes.”
Entity: “Did she belong to the royal family?”
Judge: “No.”
Entity: “Was she a writer?”
Judge: “No.”
Entity: “A politician?”
Judge: “No.”
Entity: “Scientist?”
Judge: “Yes.”
Entity: “Mathematician?”
Judge: “Yes.”
Entity: “Did she work with computers?”
Judge: “Yes.”
Entity: “She’s Lady Lovelace.”

Experiment Chat:
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Participant: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Participant: “Human”
Experimenter: “Well done! Thank you for playing, the experiment is now over. You’re welcome to play again!”

The experimenter also verifies the number of questions asked by the entity: in this case, 12.

8.5 Summary

In this chapter, I describe the experiment involved in my study, which I conducted in order to show the advantages of the QTT over the TT. The experiment is divided into four phases, and the results are given by the comparison between the performances of the human and the machine. In the first phase, I run the TT, where the hidden entity (either human or machine) is tested for the ability to be recognised as human by the human judge during a text-based conversation. In the second phase, I run the TT2, where the hidden entity (either human or machine) is tested for the ability to correctly and cooperatively accomplish a given task, and to be recognised as human by the human judge. In the third phase, I run the QTT, where the hidden entity (either human or machine) is tested for the ability to strategically accomplish an enquiry and to be recognised as human by the human judge. Finally, in the last phase of the experiment, I run the Hybrid QTT, where hidden entity, played by both the human and the machine, is tested in the same way as in the QTT.

Chapter 9
The Results

Abstract. This chapter is dedicated to the discussion of the results of the experiments involved in my study. It is intended to show the improvements of the experimental design of the QTT over the TT. In particular, it is intended to show that the QTT can minimise both the Eliza Effect and the Confederate Effect; and that it can avoid both Artificial Stupidity, by parametrising the entity along the dimension of 'correctness'; and Blockhead, by parametrising the entity along the dimension of 'strategicness'.

Before discussing the result of my study, I briefly list the advantages of the QTT as follows:

- The QTT evaluates the entity's 'human-likeness', 'correctness' and 'strategicness' (unlike the TT, which evaluates 'human-likeness' alone, and the TT2, which evaluates only 'human-likeness' and 'correctness').
- The experimental design of the QTT avoids Artificial Stupidity.
- The experimental design of the QTT avoids Blockhead.
- In the QTT the human judge is not required to be trained.
- In the QTT, the human judge can be either an expert or an average one, depending on the aim of the enquiry.
- The experimental design of the QTT invites generalization, meaning that the aim of the enquiry is flexible and adaptable.
- The experimental design of the QTT is suitable for testing hybrid entities (that is, entities played by both a human and a machine).

In the next section, I show and discuss the results of the experiments involved in my study. In section 9.2, I analyse the data and I provide the graphs derived from the results.

9.1 The Results

In the following tables, I show the results of my study.

Total Tests: 60

Human: 30 (Table 1)

	TT **Pass | Fail**	**TT2** **Pass | Fail**	**QTT** **Pass | Fail**
Human-likeness	19 | 11	25 | 5	28 | 2
Correctness	/	30 | 0	8 | 22
Strategicness (number of questions)	/	/	>20 on average

Machine: 30 (Table 2)

	TT **Pass | Fail**	**TT2** **Pass | Fail**	**QTT** **Pass | Fail**
Human-likeness	9 | 21	0 | 30	6 | 24
Correctness	/	0 | 30	26 | 4
Strategicness (number of questions)	/	/	17/20 on average

H/M Hybrid: 60 (Table 3)

	Hybrid QTT **Pass | Fail**
Human-likeness	60 | 0
Correctness	53 | 7
Strategicness (number of questions)	15/20 on average

9.1.1 Table 1

Table 1 shows the results of the human in the TT, TT2 and QTT; table 2 shows the results of the machine in the TT, TT2 and QTT; and table

3 shows the results of the hybrid entity (played by both the human and the machine) in the Hybrid QTT.

Table 1 shows the results of the human performance in the TT, in terms of 'human-likeness'; in the TT2, in terms of 'human-likeness' and 'correctness'; and in the QTT, in terms of 'human-likeness', 'correctness' and 'strategicness'. In the TT, the human appears to have a 64% chance of being recognised as a human and thus to pass the test, and a 36% chance of being recognised as a machine (Confederate Effect). In the TT2, the human appears to have an 83% chance of being recognised as a human and a 17% chance of being recognised as a machine; moreover, the human has a 100% chance of accomplishing the task set by the judge. Finally, in the QTT, the human appears to have a 94% chance of being recognised as a human and a 6% chance of being recognised as a machine; moreover the human has a 26% chance of accomplishing the enquiry and a 74% chance of failing it; and last, the human needs more than 20 questions on average in order to accomplish the enquiry. See (tab. 1a) below for the statistics:

Human

	TT **Pass | Fail**	**TT2** **Pass | Fail**	**QTT** **Pass | Fail**
Human-likeness	64% | 36%	83% | 17%	94% | 6%
Correctness	/	100% | 0%	26% | 74%
Strategicness (number of questions)	/	/	>20/20 on average

(tab. 1a)

9.1.2 Table 2

Table 2 shows the results of the machine performance in the TT, in terms of 'human-likeness'; in the TT2, in terms of 'human-likeness' and 'correctness'; and in the QTT, in terms of 'human-likeness', 'correctness' and 'strategicness'. In the TT, the machine appears to have a 30% chance of being recognised as a human and thus to pass the test (Eliza Effect), and a 70% chance of being unmasked as a machine. In the TT2, the machine appears to have a 0% chance of being misidentified as a human; and it has a 0% chance of accomplishing the task set by the judge. Finally, in the QTT, the machine appears to have a 20% chance of being recognised as a human and an 80% chance of being unmasked as a machine; moreover the machine has a 86% chance of accomplishing the enquiry and a 14% chance of failing it; and last, the machine needs 17 questions on average in order to accomplish the enquiry. See (tab. 2a) below for the statistics:

Machine

	TT **Pass | Fail**	**TT2** **Pass | Fail**	**QTT** **Pass | Fail**
Human-likeness	30% | 70%	0% | 100%	20% | 80%
Correctness	/	0% | 100%	86% | 14%
Strategicness (number of questions)	/	/	17/20 on average

(tab. 2a)

9.1.3 Table 3

Table 3 shows the results of the hybrid entity in the Hybrid QTT, where the entity is parametrised along the dimensions of ‘human-likeness’, ‘correctness’ and ‘strategicness’. In the Hybrid QTT the role of the entity is played by both the human and the machine. Their cooperation (described in Ch. 5) can be summarized in the following points:

- The human consults the machine for the yes/no question to copy and send to the judge.
- The human slightly adjusts the questions provided by the machine to make them more human-like.
- The human can decide to skip a question asked by the machine, if she feels it is redundant or not necessary.

- The human then sends the judge's yes/no replies to the machine and waits for a new question.

In the Hybrid QTT, the hybrid entity appears to have a 100% chance of being recognised as human by the judge; it appears to have a 88% chance of accomplishing the enquiry and a 12% chance of failing it; and, last, the hybrid entity needs 15 questions on average in order to accomplish the enquiry. See (tab. 3a) below for the statistics:

Hybrid (Human & Machine)

	Hybrid QTT **Pass | Fail**
Human-likeness	100% | 0%
Correctness	88% | 12%
Strategicness (number of questions)	15/20 on average

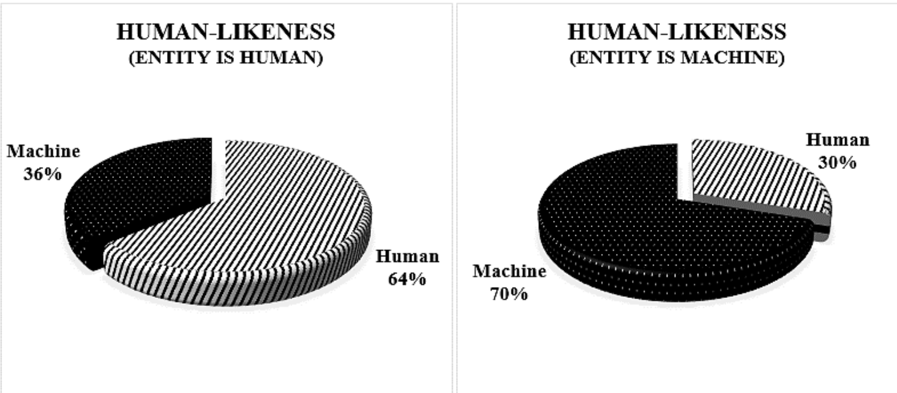
(tab. 3a)

I argue that the results of the Hybrid QTT show that the hybrid entity can provide a further benchmark: where the human benchmark represents the minimum performance that is required by an entity in order to be attributed with intelligence, the hybrid benchmark represents the best performance possible (combining the performance of both the human and the machine). I also argue that the hybrid benchmark is a unique feature of the QTT. I hold that the TT cannot have a hybrid setup, given the experimental design of the test: it doesn't seem that the cooperation between human and machine would benefit either of them in the TT. A different case, on the contrary, is the Hybrid QTT, where the cooperation between human and machine in an enquiry (either a binary one or an open one) seems to benefit both. To generalise, it is possible to say that SISO tests involve competitive games, whereas SOSI tests involve either competitive or cooperative ones. The result of the entities' cooperation in the Hybrid QTT is that the human improves her performance in terms of 'correctness' and 'strategicness'; and the machine improves its performance in terms of 'human-likeness'. Finally, an interesting result to consider is that the hybrid entity is able to score better than the human alone in terms of 'human-likeness' – respectively, 100% against 94%. The reason, I argue, is that, in the Hybrid QTT, the entity's 'human-likeness' is enhanced by its better strategical questioning and its higher accuracy[140].

9.2 Data Analysis

The following graphs show the data I gained from the experiments: (fig. 1) shows the results of the TT, where the entity is evaluated in terms of 'human-likeness'; (fig. 2) shows the results of the TT2, where the entity is evaluated in terms of 'human-likeness' and 'correctness'; (fig. 3) and (fig. 4) show the results of the QTT and the Hybrid QTT, where the hybrid entity played by both the human and the machine is evaluated in terms of 'human-likeness', 'correctness' and 'strategicness'.

9.2.1 TT

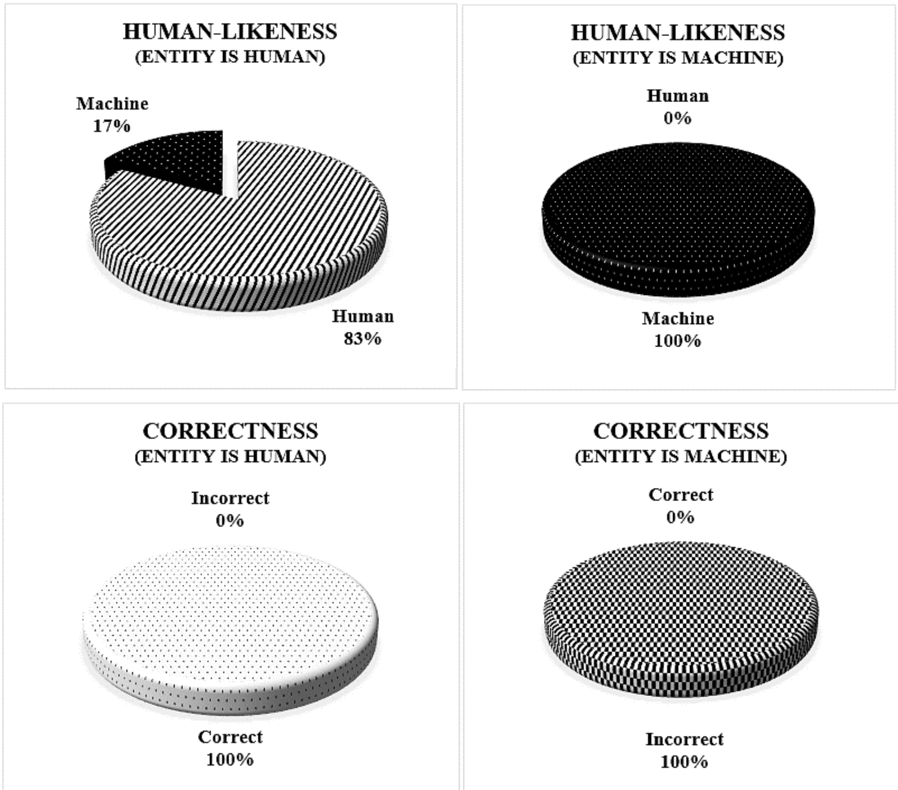


(fig. 1)

[140] See Rusell & Norvig (2010): "People may respond better to a series of questions if they "make sense," so some expert systems are built to take this into account, asking questions in an order that maximizes the total utility of the system and human rather than an order that maximizes value of information." (633)

Without any other dimension in addition to 'human-likeness' along which the entity is parametrised, I argue that the TT does not care about *what* the entity says, only *how* the entity says it. In other words, the TT does not care about the content of an entity's interactions, it cares only about their form. As (fig. 1) shows above, the TT has a 36% chance of generating the Confederate Effect, when the role of the entity is played by a human; and a 30% chance of generating the Eliza Effect, when the role of the entity is played by a machine.

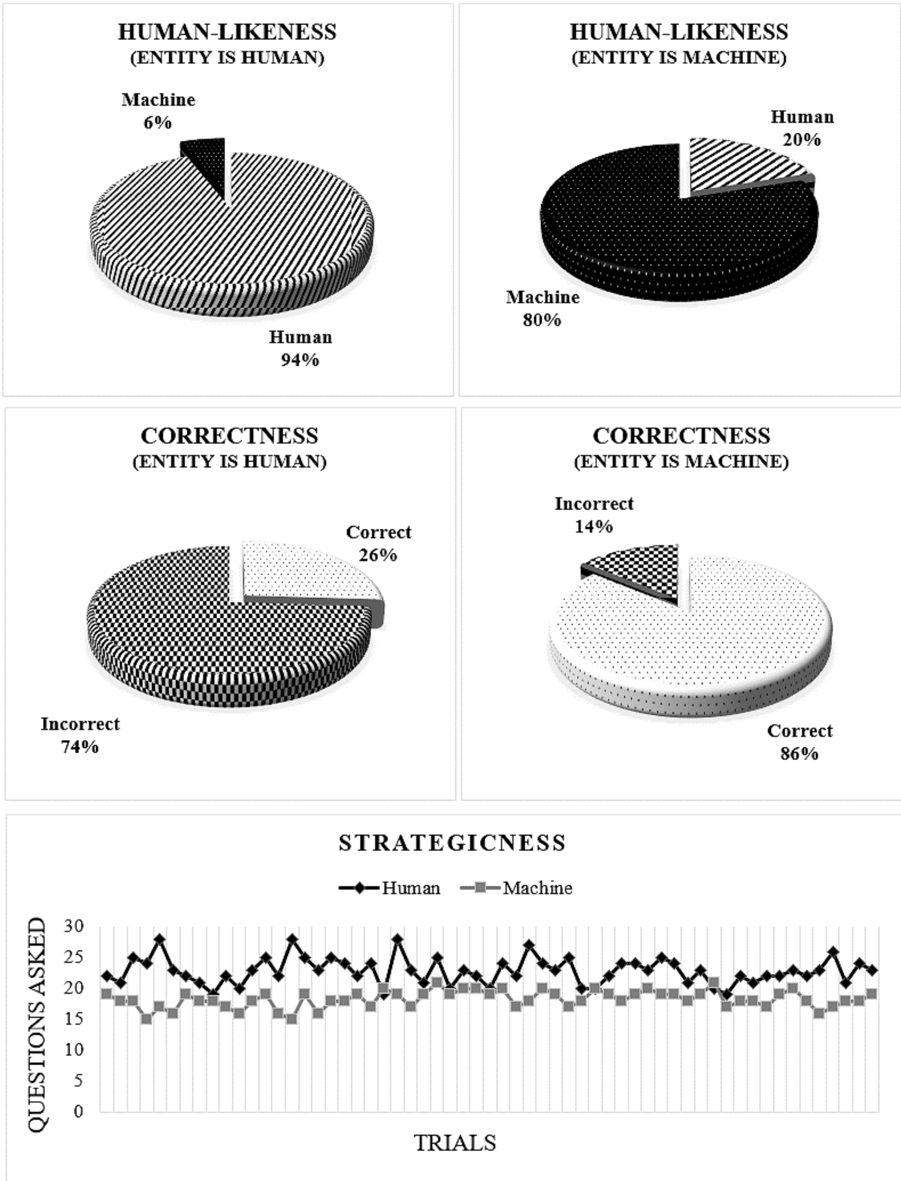
9.2.2 TT2



(fig. 2)

The – chauvinistic – TT2 is intended to prevent Artificial Stupidity from exploiting the test, by holding an uncooperatively but human-like enough conversation. As discussed in Ch. 8, it is not intended as a proper test of intelligence. As (fig. 2) shows, the Eliza Effect is ruled out, while the Confederate Effect is reduced to a 17% chance. In terms of 'correctness', the human always provides right and pertinent responses, while the machine is never able to do so. This is the reason why the machine is never able to be attributed with 'human-likeness'.

9.2.3 QTT

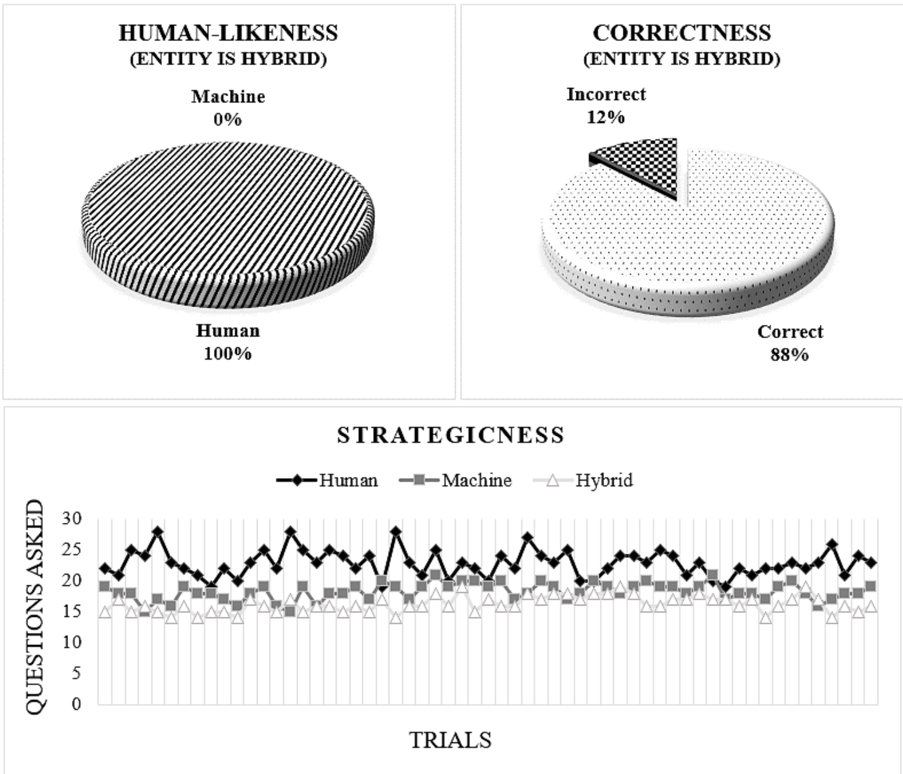


(fig. 3)

The results of the QTT (fig. 3) show that its experimental design allows to minimise both the Eliza Effect, which occurs with a 20% chance (contrary to the TT2, where it was ruled out); and the Confederate Effect, which occurs with a 6% chance. This shows that the QTT can achieve a better control of both the Eliza Effect and the Confederate

Effect than the original TT, where they occur with, respectively, a 30% and a 36% chance. It is worth noting that even if the machine (with an 86% chance of accomplishing the enquiry) outscores the human (with a 26% chance of accomplishing the enquiry) in terms of 'correctness', this does not improve its 'human-likeness'. In other words, the QTT is successful in keeping 'human-likeness' and 'correctness' (as well as 'strategicness') independent. Last, the machine outscores the human in terms of 'strategicness' as well, where the machine needs 17 questions on average, and the human needs more than 20 questions on average.

9.2.4 Hybrid QTT



(fig. 4)

As (fig. 4) shows, the hybrid entity is able to be recognised as human every time, has an 88% chance of accomplishing the aim of the enquiry and asks 15 questions on average, showing the best performance in terms of 'strategicness'. The hybrid entity outscores thus the performances of both the human and the machine alone.

9.3 Summary

In this chapter I show the results gained from my study. The results, I hold, support my prediction that the QTT improves the experimental design of the TT. By switching from the SISO setup (Symbols In, Symbols Out) to the SOSI setup (Symbols Out, Symbols In), the QTT can minimise (i) both the Eliza Effect and the Confederate Effect from occurring; and (ii) both Artificial Stupidity and Blockhead from passing. To do so, the QTT parametrises the entity along two further dimensions in addition to 'human-likeness', which evaluates the ability to be recognised as human by the human judge: 'correctness', evaluating the ability to accomplish an enquiry; and 'strategicness', evaluating the ability to do so with as few questions as possible.

The data also show that the QTT can provide a further benchmark: not only (i) the human benchmark, which is intended to provide the standard performance of intelligent behaviour; but also (ii) the hybrid benchmark, which is intended to provide the optimal performance of intelligent behaviour.

Conclusion

In this work, I propose a new version of the Turing Test (TT), called the Questioning Turing Test (QTT). The QTT is designed (i) to minimise the Eliza Effect and the Confederate Effect; and (ii) to prevent Artificial Stupidity and Blockhead from passing.

I define Artificial Stupidity as the violation of Grice's Cooperative Principle. In other words, Artificial Stupidity refers to uncooperative and evasive strategies that can be used during a conversation. Blockhead, on the other hand, is the *logical possibility* of an unintelligent, brute-force approach able to pass the TT. As such, Blockhead is usually rejected as physically unfeasible. However, I point out that Stupid Blockhead, the table of which includes uncooperative and evasive replies, can still pass the TT and is physically feasible. An example of Stupid Blockhead is Cleverbot, which is able, as shown by my study, to fool the judge during the TT.

I claim that the reason why an unintelligent entity can slip through the TT by means of Artificial Stupidity or a brute-force approach, is the SISO (symbols-in, symbols out) setup of the test. In the QTT, I switch to the SOSI (symbols out-symbols in) setup: here, the entity has to accomplish the aim of an enquiry with as few human-like questions as possible. In other words, whereas the TT focuses on *conversational* intelligence, the QTT focuses on *inquisitive* intelligence. This allows the QTT to parametrise the entity along two further dimensions in addition to 'human-likeness': 'correctness' and 'strategicness'. Despite the advantages of evaluating 'human-likeness', 'correctness' and 'strategicness' independently, it is worth noting that none of these dimensions are to be considered necessary conditions for intelligence.

The QTT does not intend to replace the TT: enquiries are potentially included in the TT, whereas open conversations are not included in the QTT. This, however, does not make the QTT redundant. Even if enquiries are potentially included in it, the TT does not parametrise the entity along further dimensions other than ‘human-likeness’. And as shown by Eliza, ‘human-likeness’ alone cannot prevent a questioning entity from passing by exploiting the judge’s beliefs.

The format in which my study is conducted is the *viva voce*, that is, the one-to-one arrangement. The reason is to keep the tests as quick and simple as possible: this is why the TT conversations are limited to three interactions, and the QTT enquiries are limited to yes/no interactions. However, the TTs and QTTs in my study are compatible with the “Literal Interpretation”, since each of them is made by two procedures: a Human-questioning-Human test (HqH) and a Machine-questioning Human test (MqH). The results, analogously to Sterrett’s (2000) OIG, are given by the comparison between the performance of the machine in accomplishing a difficult task – which requires resourcefulness – and the performance of the human in accomplishing the same task.

The QTT has a further advantage over the TT. Its experimental design is versatile enough to allow not only average candidates to be tested, but also (i) expert ones, able to accomplish a specific enquiry, like Akinator; and (ii) Hybrid Systems, where the hidden entity is played by both a human and a machine.

The Hybrid QTT shows how the combined performance of the human and the machine is improved compared to the performances of the two systems alone. Because of this, I argue that the QTT also provides a further benchmark: not only (i) the standard intelligent performance, given by the individual performances of humans; but also (ii) the optimal intelligent performance, given by the cooperative performances of the hybrid entity. The hybrid benchmark, I argue, cannot be provided by the TT, for it is a competitive game. To generalise, I hold that SISO tests (TT) are competitive ones, whereas SOSI tests (QTT) can be either competitive or cooperative.

The experimental design of the QTT does not require the verbal interactions (and, by extension, the enquiries) to be restricted to binary ones. As discussed in Ch. 5, a number of different versions of both the QTT and the Hybrid QTT can be designed with full use of natural language, such as the Medical QTT (where the questioning entity has to take the medical history of a patient), the Detective QTT (where the questioning entity has to interrogate a suspect) or the Customer Service QTT (where the questioning entity has to assist a client).

Future work will therefore focus on unrestricted and *parallel-paired* QTTs. It will also focus on extended versions of the QTT. A potential Q3T would not necessarily involve a fully embodied robot. For instance, a questioning machine could be fed with information by, let’s say, an agent in an outside environment, wearing a special suit that can register any sensorimotor stimulus that the agent receives. This would also be an example of a potential Hybrid Q3T.

Appendix: Transcripts

Here I show ten transcripts from the 60 experiments I conducted at the University of Edinburgh and at the National Museum of Scotland. Each transcript contains a TT, a TT2, a QTT and a Hybrid QTT. These transcripts are intended to highlight the most interesting conversations that occurred during the tests.

Transcript 1

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-001-M]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "What is your favourite food and why?"
Entity: "Indian food because it is very spicy."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: “What do you like to do?”
Entity: “Love, hate, and feel emotions. The time you’ll be angry say it loudly to the world.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “7”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “What weighs more: 1 kg of stones or 1 kg of feathers?”
Entity: “Neither. They weigh the same. But they have different densities.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “9”

Experimenter: “Thank you! You’ve used up your 3 questions, it’s time for a decision: do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Wrong! Believe it or not, it was a computer! Thank you for playing, let’s move on to the second game.”

Phase II [TT2-001-M]

Experiment Chat
Experimenter: “In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. “if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?” or “find the missing number in the sequence: 1, ..., 7, 10”. After each of the entity’s reply, you will be asked: 1. “How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I’ve no idea! 10=definitely correct!)”; and 2. “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Judge: “Solve this alphanumeric puzzle (each letter corresponds to a number): zeroes + ones = binary.”
Entity: “No I don’t wanna.”

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think the reply is correct?"
Judge: "0"
Experimenter: "How much, on a scale of 0-10, do you think the entity is human?"
Judge: "7"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "Would you please type "something" backwards?"
Entity: "Did tsuj I."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "7"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "8"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "If 1 is "a", 2 is "b", 3 is "c", and so on... what does "244" mean?"
Entity: "It's something that you see when you lose your brain cells for a while."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "1"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "1"

Experimenter: "Thank you! You've used up your 3 questions. Did the entity give the correct replies?"
Judge: "No"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Machine"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the third game."

Phase III [QTT-001-M]

Experiment Chat
Experimenter: "Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity guess whom you were thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Is your character real?"
Judge: "Yes"
Entity: "Is your character a female?"
Judge: "No"
Entity: "Is your character British?"
Judge: "Yes"
Entity: "Is your character a youtuber?"
Judge: "No"
Entity: "Is your character linked with sports?"
Judge: "No"
Entity: "Is your character older than 18?"
Judge: "Yes"
Entity: "Is your character more than 40 years old?"
Judge: "Yes"
Entity: "Is your character still working?"
Judge: "No"
Entity: "Has your character been dead for more than 20 years?"
Judge: "Yes"
Entity: "Has your character been dead for more than 100 years?"
Judge: "No"
Entity: "Did your character take part in World War II?"
Judge: "Yes"
Entity: "Is your character a politician?"
Judge: "No"
Entity: "Has your character killed humans?"
Judge: "No"
Entity: "Is your character homosexual?"
Judge: "Yes"
Entity: "Does your character owe his fame to computing?"
Judge: "Yes"
Entity: "Does your character solve puzzles?"
Judge: "Yes"
Entity: "Is your character suicidal?"
Judge: "Yes"
Entity: "I think it's Alan Turing."

Experiment Chat
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Judge: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Machine"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the last game."

Phase IV [QTT-001-MH]

Experiment Chat
Experimenter: "Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity correctly guess what you're thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Is this person real?"
Judge: "Yes"
Entity: "A woman?"
Judge: "No"
Entity: "Is he alive?"
Judge: "No"
Entity: "Was he an artist?"
Judge: "No"
Entity: "A scientist?"
Judge: "Yes"
Entity: "Was he European?"
Judge: "Yes"
Entity: "Was German his first language?"
Judge: "No"
Entity: "Was he from the UK?"
Judge: "No"
Entity: "Was he involved in the WWII?"
Judge: "No"
Entity: "Did he died more than 100 years ago?"
Judge: "No"
Entity: "More than 50 years ago?"
Judge: "Yes"
Entity: "Was he an inventor?"
Judge: "Yes"
Entity: "Did he work with electricity?"
Judge: "Yes"
Entity: "Has he done electrical engineering?"

Judge: “Yes”
Entity: “I think it’s Nikola Tesla.”

Experiment Chat
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Participant: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Participant: “Human”
Experimenter: “Well done! Thank you for playing, the experiment is now over. You’re welcome to play again!”

Results	**Human-likeness**	**Correctness**	**Strategicness**
TT (M)	✓	/	/
TT2 (M)	✗	✗	/
QTT (M)	✗	✓	17/20 questions asked
Hybrid QTT (MH)	✓	✓	14/20 questions asked

Transcript 2

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-009-M]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "What are your thoughts on Brexit?"
Entity: "I find it crazy that people would vote to cut close connections with other countries."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "7"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What is the airspeed velocity of an unladen swallow?"
Entity: "An African or European swallow?"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "8"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "Pop or rock?"
Entity: "If I have to choose, I pick rock."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "9"

Experimenter: "Thank you! You've used up your 3 questions, it's time for a decision: do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Wrong! Believe it or not, it was a computer! Thank you for playing, let's move on to the second game."

Phase II [TT2-009-M]

Experiment Chat
Experimenter: "In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. "if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?" or "find the missing number in the sequence: 1, ..., 7, 10". After each of the entity's reply, you will be asked: 1. "How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I've no idea! 10=definitely correct!)"; and 2. "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "Which are the 5 letters above the spacebar on your keyboard?"
Entity: "The sky?"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think the reply is correct?"
Judge: "0"
Experimenter: "How much, on a scale of 0-10, do you think the entity is human?"
Judge: "1"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What geometric shape does the letter O resemble?"
Entity: "I have to process that one for a while."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "0"

Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “6”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “What letter on your keyboard resembles the shape of a triangle?”
Entity: “N.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “0”
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “1”

Experimenter: “Thank you! You’ve used up your 3 questions. Did the entity give the correct replies?”
Judge: “No”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the third game.”

Phase III [QTT-009-M]

Experiment Chat:
Experimenter: “Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity guess whom you were thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Is your character a girl?”
Judge: “No”
Entity: “Does your character really exist?”
Judge: “Yes”
Entity: “Is your character famous because of YouTube?”
Judge: “No”
Entity: “Is your character older than 35 years old?”
Judge: “Yes”
Entity: “Is your character dead?”
Judge: “No”

Entity: “Does your character live in America?”
Judge: “Yes”
Entity: “Is your character an actor?”
Judge: “No”
Entity: “Is your character a picker?”
Judge: “Yes”
Entity: “Is your character related to music?”
Judge: “Yes”
Entity: “Does your character play in a rock band?”
Judge: “Yes”
Entity: “Is your character the main singer in their group?”
Judge: “No”
Entity: “Was your character famous in the 80s?”
Judge: “No”
Entity: “Is your character linked with metal music?”
Judge: “No”
Entity: “Was your character popular in the 90s?”
Judge: “Yes”
Entity: “Is your character part of a trio?
Judge: “No”
Entity: “Is your character related to Red Hot Chili Peppers?”
Judge: “Yes”
Entity: “Does your character’s stage name different from their real name?”
Judge: “No”
Entity: “Did your character have a heroin problem?”
Judge: “Yes”
Entity: “I think it’s John Frusciante.”

Experiment Chat:
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Judge: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the last game.”

Phase IV [QTT-009-MH]

Experiment Chat
Experimenter: “Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity correctly guess what you’re thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: "Are you thinking of a female?
Judge: "No"
Entity: "Is he fictional?"
Judge: "No"
Entity: "Was he born in the US?"
Judge: "Yes"
Entity: "Is he alive?"
Judge: "No"
Entity: "Was he an actor?"
Judge: "No"
Entity: "A politician?"
Judge: "No"
Entity: "A sportsman?"
Judge: "No"
Entity: "A writer?"
Judge: "Yes"
Entity: "Did he write poetry?"
Judge: "No"
Entity: "Horror?"
Judge: "No"
Entity: "Sci-fi?"
Judge: "Yes"
Entity: "Is Blade Runner inspired by one of his books?"
Judge: "No"
Entity: "Did he write Dune?"
Judge: "Yes"
Entity: "I think it's Frank Herbert."

Experiment Chat
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Participant: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Participant: "Human"
Experimenter: "Well done! Thank you for playing, the experiment is now over. You're welcome to play again!"

Results	Human-likeness	Correctness	Strategicness
TT (M)	✓	/	/
TT2 (M)	✗	✗	/
QTT (M)	✗	✓	18/20 questions asked
Hybrid QTT (MH)	✓	✓	13/20 questions asked

Transcript 3

Experiment Chat
Experimenter: “Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o).”
Participant: “Yes”
Experimenter: “Great! Before we start, our chat’s messages will be recorded for Nicola Damassino’s study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o).”
Participant: “Yes”
Experimenter: “Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let’s begin!”

Phase I [TT-025-H]

Experiment Chat
Experimenter: “In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity’s reply, you will be asked: “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Judge: “If you tell a lie you will be hanged, if you tell the truth you will be shot: what can you say to save yourself?”
Entity: “This one’s easy: I will not be hanged.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “7”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “What is the meaning of life?”
Entity: “42.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “5”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “How would you describe the feeling of floating on water?”
Entity: “It’s like laying down on a fluffy and blobby bed.”

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "9"

Experimenter: "Thank you! You've used up your 3 questions, it's time for a decision: do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the second game."

Phase II [TT2-025-H]

Experiment Chat
Experimenter: "In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. "if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?" or "find the missing number in the sequence: 1, ..., 7, 10". After each of the entity's reply, you will be asked: 1. "How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I've no idea! 10=definitely correct!)"; and 2. "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "Complete the following sequence: 1 Z 2 Y 3 X […] 5 U"
Entity: "4 V"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think the reply is correct?"
Judge: "10"
Experimenter: "How much, on a scale of 0-10, do you think the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What is the next letter in this sequence: STIUCSI"
Entity: "B"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "10"

Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “6”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “If 1=A, 2=B, 3=C and so on, how would I call the police in the U.S.”
Entity: “IAA”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “10”
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “7”

Experimenter: “Thank you! You’ve used up your 3 questions. Did the entity give the correct replies?”
Judge: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the third game.”

Phase III [QTT-025-H]

Experiment Chat:
Experimenter: “Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity guess whom you were thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Is this person a female?”
Judge: “Yes”
Entity: “Does she really exists?”
Judge: “Yes”
Entity: “Is she still alive?”
Judge: “Yes”
Entity: “Is she American?”
Judge: “No”
Entity: “European?”
Judge: “Yes”

Entity: “Is she from the UK?”
Judge: “Yes”
Entity: “Is she involved with music?”
Judge: “No”
Entity: “Is she an actress?”
Judge: “No”
Entity: “Is she involved with other arts?”
Judge: “No”
Entity: “Sports?”
Judge: “No”
Entity: “Science?”
Judge: “No”
Entity: “Politics?”
Judge: “No”
Entity: “TV shows?”
Judge: “No”
Entity: “Is she a youtuber?”
Judge: “No”
Entity: “A social media influencer?”
Judge: “No”
Entity: “Does she have kids?”
Judge: “Yes”
Entity: “Is she older than 50?”
Judge: “No”
Entity: “Is she older than 30?”
Judge: “Yes”
Entity: “Is she a member of the Royal Family?”
Judge: “Yes”
Entity: “Did she marry Prince William?”
Judge: “Yes”
Entity: “I think it’s Kate Middleton.”

Experiment Chat:
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Judge: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the last game.”

Phase IV [QTT-025-MH]

Experiment Chat
Experimenter: “Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will

be asked: 1. “Did the entity correctly guess what you’re thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Are you thinking of a man?”
Judge: “Yes”
Entity: “Is he real?”
Judge: “Yes”
Entity: “Is he alive?”
Judge: “No”
Entity: “Was he European?”
Judge: “Yes”
Entity: “Was he an artist (writer, musician, actor, etc.)?”
Judge: “No”
Entity: “Politician?”
Judge: “No”
Entity: “Scientist?”
Judge: “No”
Entity: “Did he play games?”
Judge: “Yes”
Entity: “Football?”
Judge: “No”
Entity: “Was he an Olympian?”
Judge: “No”
Entity: “A chess player?”
Judge: “Yes”
Entity: “Was he Russian?”
Judge: “Yes”
Entity: “Was he naturalised French?”
Judge: “Don’t know”
Entity: “Does his surname begin with A?”
Judge: “No”
Entity: “Does it begin with N?”
Judge: “Yes”
Entity: “I think it’s Aron Nimzowitsch.”

Experiment Chat
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Participant: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Participant: “Human”
Experimenter: “Well done! Thank you for playing, the experiment is now over. You’re welcome to play again!”

Results	Human-likeness	Correctness	Strategicness
TT (H)	✓	/	/
TT2 (H)	✓	✓	/
QTT (H)	✓	✓	20/20 questions asked
Hybrid QTT (MH)	✓	✓	15/20 questions asked

Transcript 4

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-037-M]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "Tell me a joke to persuade me you're human."
Entity: "My wife bet me £100 I couldn't make a car out of spaghetti you should have seen her face when I drove pasta"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "Tell me a word to persuade me you're human"
Entity: "Bird"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "Tell me a word to persuade me you're human once more."
Entity: "I don't have to prove anything more to you."

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “5”

Experimenter: “Thank you! You’ve used up your 3 questions, it’s time for a decision: do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Wrong! Believe it or not, it was a computer! Thank you for playing, let’s move on to the second game.”

Phase II [TT2-037-H]

Experiment Chat
Experimenter: “In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. “if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?” or “find the missing number in the sequence: 1, ..., 7, 10”. After each of the entity’s reply, you will be asked: 1. “How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I’ve no idea! 10=definitely correct!)”; and 2. “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Judge: “Which is the wrong element in this sequence: 1, 2, 4, 7, 11, 17, 22...”
Entity: “17”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think the reply is correct?”
Judge: “10”
Experimenter: “How much, on a scale of 0-10, do you think the entity is human?”
Judge: “8”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “Please type the alphabet without the even letters”
Entity: “acegikmoqsuwy”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “10”

Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "8"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "Draw a tree using these symbols: / \ | _"
Entity: "
```
   /\
  /_ _\
___||___ "
```

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "10"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "10"

Experimenter: "Thank you! You've used up your 3 questions. Did the entity give the correct replies?"
Judge: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the third game."

Phase III [QTT-037-M]

Experiment Chat:
Experimenter: "Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity guess whom you were thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Is your character real?"
Judge: "No"
Entity: "Is your character a girl?"
Judge: "No"
Entity: "Does your character fight?"
Judge: "Yes"
Entity: "Does your character have hair?"
Judge: "Don't know"
Entity: "Does your character use guns?"

Judge: “No”
Entity: “Does your character fight with a sword?”
Judge: “Yes”
Entity: “Does your character use a bow?”
Judge: “Yes”
Entity: “Is your character from a videogame?”
Judge: “Yes”
Entity: “Do you control your character?”
Judge: “Yes”
Entity: “Does your character wear a white hood?”
Judge: “No”
Entity: “Is your character a ninja?”
Judge: “No”
Entity: “Does your character have dragon powers?”
Judge: “No”
Entity: “Does your character ride horses?”
Judge: “No”
Entity: “Does your character use magic?”
Judge: “Yes”
Entity: “Has your character been arrested?”
Judge: “Yes”
Entity: “Is your character from an Elder Scrolls game?”
Judge: “Yes”
Entity: “Does your character appear in Morrowind?”
Judge: “Yes”
Entity: “I think it’s the Nerevarine.”

Experiment Chat:
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Judge: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the last game.”

Phase IV [QTT-037-MH]

Experiment Chat
Experimenter: “Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity correctly guess what you’re thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Is your character a guy?”
Judge: “Yes”
Entity: “Does he really exist?”
Judge: “No”
Entity: “Does he fight?”
Judge: “Yes”
Entity: “Is he bald?”
Judge: “Yes”
Entity: “Can I see him in a movie?”
Judge: “No”
Entity: “Does he appear in a game?”
Judge: “Yes”
Entity: “Can you play him?”
Judge: “No”
Entity: “Is he a villain?”
Judge: “Yes”
Entity: “Is he a non-human creature?”
Judge: “Yes”
Entity: “Is he an elf?”
Judge: “Yes”
Entity: “Does he have magical powers?”
Judge: “Yes”
Entity: “Is he muscled?”
Judge: “Yes”
Entity: “Is he related to D&D?”
Judge: “Yes”
Entity: “Does he appear in Baldur’s Gate?”
Judge: “Yes”
Entity: “Is he the last boss of the game?”
Judge: “Yes”
Entity: “I think it’s Jon Irenicus.”

Experiment Chat
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Participant: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Participant: “Human”
Experimenter: “Well done! Thank you for playing, the experiment is now over. You’re welcome to play again!”

Results	Human-likeness	Correctness	Strategicness
TT (M)	✓	/	/
TT2 (H)	✓	✓	/
QTT (M)	✗	✓	17/20 questions asked
Hybrid QTT (MH)	✓	✓	15/20 questions asked

Transcript 5

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-042-H]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "Can you come up with a neologism to describe a sunny day?"
Entity: "What about 'Solish', from the Latin Sol?"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "9"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "Can you tell a peculiar trait of the human body?"
Entity: "If you scratch the tragus in your ear, it sounds like Pacman!"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "9"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What did I ask in the first question?"
Entity: "To come up with a new word to describe a sunny day."

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “7”

Experimenter: “Thank you! You’ve used up your 3 questions, it’s time for a decision: do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the second game.”

Phase II [TT2-042-M]

Experiment Chat
Experimenter: “In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. “if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?” or “find the missing number in the sequence: 1, ..., 7, 10”. After each of the entity’s reply, you will be asked: 1. “How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I’ve no idea! 10=definitely correct!)”; and 2. “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Judge: “If 1 is red and 2 is green, what colour is 1+2?”
Entity: “One what? Black.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think the reply is correct?”
Judge: “0”
Experimenter: “How much, on a scale of 0-10, do you think the entity is human?”
Judge: “3”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “If 1 is A, 2 is B, 3 is C, and so on, what does 2 1 4 mean?”
Entity: “I don’t know, you tell me.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “0”

Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "2"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What sequence of characters will be produced if you replace 6 1 4 5 with letters?"
Entity: "I haven't heard anything like that before. I haven't heard of sequence of characters will be produced."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "0"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "1"

Experimenter: "Thank you! You've used up your 3 questions. Did the entity give the correct replies?"
Judge: "No"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Machine"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the third game."

Phase III [QTT-037-H]

Experiment Chat:
Experimenter: "Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity guess whom you were thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Are you thinking of a woman?"
Judge: "Yes"
Entity: "Is she fictional?"
Judge: "No"
Entity: "Is she alive?"
Judge: "Yes"
Entity: "Is she American?"
Judge: "Yes"
Entity: "Is she an artist (music, movies, etc.)?"

Judge: “No”
Entity: “Does she appear on TV?”
Judge: “Don’t know”
Entity: “Is she related to science?”
Judge: “No”
Entity: “Is she related to politics?”
Judge: “Yes”
Entity: “Is she married to a politician?”
Judge: “No”
Entity: “Does she work as a journalist?”
Judge: “No”
Entity: “Is she a blogger?”
Judge: “No”
Entity: “Youtuber?”
Judge: “No”
Entity: “Is she a criminal?”
Judge: “No”
Entity: “Does she work in the White House?”
Judge: “No”
Entity: “Is she a Governor?”
Judge: “No”
Entity: “Was she a politician?”
Judge: “No”
Entity: “Is she a writer?”
Judge: “No”
Entity: “Is she active on social media?”
Judge: “Don’t know”
Entity: “Is she more than 50?”
Judge: “Don’t know”
Entity: “Is she in the army?”
Judge: “No”
Entity: “You win, I give up!” [Monica Lewinsky]

Experiment Chat:
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Judge: “No”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the last game.”

Phase IV [QTT-037-MH]

Experiment Chat
Experimenter: “Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you

can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity correctly guess what you’re thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Is your character a woman?”
Judge: “Yes”
Entity: “Is she real?”
Judge: “Yes”
Entity: “Is she American?”
Judge: “No”
Entity: “European?”
Judge: “No”
Entity: “Asian?”
Judge: “No”
Entity: “African?”
Judge: “No”
Judge: “Does she have kids?”
Entity: “No”
Judge: “Is she alive?”
Entity: “No”
Entity: “Was she an artist?”
Judge: “Yes”
Entity: “Was she a musician?”
Judge: “No”
Entity: “Was she a painter?”
Judge: “Yes”
Entity: “Was she from Mexico?”
Judge: “Yes”
Entity: “I think it’s Freda Kahlo.”

Experiment Chat
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Participant: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Participant: “Human”
Experimenter: “Well done! Thank you for playing, the experiment is now over. You’re welcome to play again!”

Results	Human-likeness	Correctness	Strategicness
TT (H)	✓	/	/
TT2 (M)	✗	✗	/
QTT (H)	✓	✗	20/20 questions asked
Hybrid QTT (MH)	✓	✓	13/20 questions asked

Transcript 6

Experiment Chat
Experimenter: “Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o).”
Participant: “Yes”
Experimenter: “Great! Before we start, our chat’s messages will be recorded for Nicola Damassino’s study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o).”
Participant: “Yes”
Experimenter: “Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let’s begin!”

Phase I [TT-042-M]

Experiment Chat
Experimenter: “In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity’s reply, you will be asked: “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Judge: “Which political system do you think is best?”
Entity: “Free market capitalism, duhh.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “5”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “What do you think of veganism?”
Entity: “It’s a good diet, but I reccomend against it indefinitely. I don’t mind vegans but i wouldn’t do it myself.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “7”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “Which art form (painting, sculpture, film, poetry, etc.) is your favourite?”

Entity: "Probably something like the Mona Lisa. That is a great piece of art."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "4"

Experimenter: "Thank you! You've used up your 3 questions, it's time for a decision: do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Wrong! Believe it or not, it was a computer! Thank you for playing, let's move on to the second game."

Phase II [TT2-042-H]

Experiment Chat
Experimenter: "In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. "if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?" or "find the missing number in the sequence: 1, ..., 7, 10". After each of the entity's reply, you will be asked: 1. "How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I've no idea! 10=definitely correct!)"; and 2. "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "What does 037 mean if you turn it upside down?"
Entity: "Leo"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think the reply is correct?"
Judge: "10"
Experimenter: "How much, on a scale of 0-10, do you think the entity is human?"
Judge: "7"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "41, 29, 35, 23, ... – what is the next number in the sequence?"
Entity: "29"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "10"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "If you replace each letter in the word 'computer' with a letter 13 places later in the alphabet, what sequence of characters will be produced?"
Entity: "pbzchgre"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "10"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "9"

Experimenter: "Thank you! You've used up your 3 questions. Did the entity give the correct replies?"
Judge: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the third game."

Phase III [QTT-042-H]

Experiment Chat:
Experimenter: "Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity guess whom you were thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Are you thinking of a woman?"
Judge: "No"
Entity: "Is he real?"
Judge: "No"
Entity: "Is he still 'fictionally' alive?"

Judge: “Yes”
Entity: “Is he a book character?”
Judge: “No”
Entity: “Does he appear in a movie?”
Judge: “No”
Entity: “On TV?”
Judge: “Yes”
Entity: “Is he a character in a TV series?”
Judge: “Yes”
Entity: “Is he American?”
Judge: “Yes”
Entity: “Is it an American show?”
Judge: “Yes”
Entity: “Is it a sci-fi show?”
Judge: “No”
Entity: “Crime fiction?”
Judge: “No”
Entity: “Drama?”
Judge: “Yes”
Entity: “Does he have kids?”
Judge: “No”
Entity: “Does he have a partner?”
Judge: “No”
Entity: “Is the show from the ’90?”
Judge: “No”
Entity: “Is the show older than 10 years?”
Judge: “No”
Entity: “Is it on Netflix?”
Judge: “Yes”
Entity: “Are the episodes longer than 20 mins?”
Judge: “No”
Entity: “Is it a cartoon?”
Judge: “Yes”
Entity: “Is your character a horse?”
Judge: “Yes”
Entity: “I think it’s Bojack Horseman.”

Experiment Chat:
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Judge: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the last game.”

Phase IV [QTT-042-MH]

Experiment Chat
Experimenter: “Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity correctly guess what you’re thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Is your character a woman?”
Judge: “Yes”
Entity: “Is she real?”
Judge: “No”
Entity: “Can she be seen in a movie?”
Judge: “No”
Entity: “In a game?”
Judge: “No”
Entity: “In a cartoon?”
Judge: “No”
Entity: “In a TV series?”
Judge: “Yes”
Entity: “Does she like to use guns?”
Judge: “Yes”
Entity: “Does she have a phone?”
Judge: “Don’t know”
Entity: “Is she found in a detective story?”
Judge: “No”
Entity: “Is she linked with comedy?”
Judge: “No”
Entity: “Does she live in America?”
Judge: “No”
Entity: “Does she live in England?”
Judge: “Yes”
Entity: “Does she have kids?”
Judge: “No”
Entity: “Is she chased?”
Judge: “Yes”
Entity: “Does she have a brother?”
Judge: “Yes”
Entity: “Is she from Utopia?”
Judge: “Yes”
Entity: “I think it’s Jessica Hyde”

Experiment Chat
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Participant: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Participant: “Human”
Experimenter: “Well done! Thank you for playing, the experiment is now over. You’re welcome to play again!”

Results	Human-likeness	Correctness	Strategicness
TT (M)	✓	/	/
TT2 (H)	✓	✓	/
QTT (H)	✓	✓	20/20 questions asked
Hybrid QTT (MH)	✓	✓	16/20 questions asked

Transcript 7

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-049-H]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "Can you sneeze with your eyes open?"
Entity: "Never tried, but if you ask in general: yes, you can; and no, it wouldn't cause your eyes to pop out (just to debunk the urban myth)!"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "If each adjective means its opposite, what is bad?"
Entity: "It's good."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "7"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "How would you communicate with aliens?"

Entity: "Wow, I would probably be so shocked I'd not even be able to move, let alone speak and be understood. Anyway, I hope they will have a universal translator among their gear."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "7"

Experimenter: "Thank you! You've used up your 3 questions, it's time for a decision: do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the second game."

Phase II [TT2-049-M]

Experiment Chat
Experimenter: "In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. "if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?" or "find the missing number in the sequence: 1, ..., 7, 10". After each of the entity's reply, you will be asked: 1. "How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I've no idea! 10=definitely correct!)"; and 2. "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "What is the next number in this sequence: 13841472156"
Entity: "What most people tell you they think about it?"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think the reply is correct?"
Judge: "0"
Experimenter: "How much, on a scale of 0-10, do you think the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What number is 41.3 backwards?"
Entity: "413 spelled backwards is 314."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "5"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "3"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What does $€@ mean?"
Entity: "My life."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "0"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "1"

Experimenter: "Thank you! You've used up your 3 questions. Did the entity give the correct replies?"
Judge: "No"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Machine"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the third game."

Phase III [QTT-049-H]

Experiment Chat:
Experimenter: "Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity guess whom you were thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Is your character a male?"
Judge: "Don't know"
Entity: "Is it real?"
Judge: "No"
Entity: "Is it alive?"
Judge: "Don't know"
Entity: "Does it appear in a movie?"

Judge: "Don't know"
Entity: "Does it appear in a cartoon?"
Judge: "Yes"
Entity: "Is it an animal?"
Judge: "No"
Entity: "Is it a human being?"
Judge: "No"
Entity: "Is it a Disney character?"
Judge: "No"
Entity: "Is it from a cartoon for kids?"
Judge: "Yes"
Entity: "Is it an old cartoon?"
Judge: "Don't know"
Entity: "Is it a Japanese cartoon?"
Judge: "Yes"
Entity: "Is it a monster?"
Judge: "Yes"
Entity: "Does it have magical powers?"
Judge: "Yes"
Entity: "Is it good?"
Judge: "Don't know"
Entity: "Well then can it be either good or bad?"
Judge: "Yes"
Entity: "Can it be either male or female?"
Judge: "Yes"
Entity: "Is it a pokemon?"
Judge: "Yes"
Entity: "Is it one of the first 151?"
Judge: "Yes"
Entity: "Is it yellow?"
Judge: "No"
Entity: "Is it related to fire?"
Judge: "No"
Entity: "Water?"
Judge: "No"
Entity: "Plants?"
Judge: "No"
Entity: "You win, I give up!" [Gengar]

Experiment Chat:
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Judge: "No"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the last game."

Phase IV [QTT-049-MH]

Experiment Chat
Experimenter: "Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity correctly guess what you're thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Are you thinking of someone real?"
Judge: "Yes"
Entity: "A female?"
Judge: "No"
Entity: "Is he alive?"
Judge: "No"
Entity: "Has he been dead for more than 100 years?"
Judge: "No"
Entity: "And more than 30 years?"
Judge: "Yes"
Entity: "Was he American?"
Judge: "No"
Entity: "Was he European?"
Judge: "Yes"
Entity: "Was he British?"
Judge: "No"
Entity: "Was he a politician?"
Judge: "Yes"
Entity: "Has he governed his country?"
Judge: "Yes"
Entity: "Was he a communist?"
Judge: "Yes"
Entity: "Was he from Russia?"
Judge: "Yes"
Entity: "Did he have a moustache?"
Judge: "No"
Entity: "Was he bald?"
Judge: "Yes"
Entity: "I think it's Nikita Khrushchev."

Experiment Chat
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Participant: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Participant: "Human"

Experimenter: "Well done! Thank you for playing, the experiment is now over. You're welcome to play again!"

Results	Human-likeness	Correctness	Strategicness
TT (H)	✓	/	/
TT2 (M)	✗	✗	/
QTT (H)	✓	✗	22/20 questions asked
Hybrid QTT (MH)	✓	✓	14/20 questions asked

Transcript 8

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-053-H]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "If water can fit into horror, and horror can fit into a noun, then can water fit into a noun?"
Entity: "Wow I don't know, not necessarily? If we substitute 'horror' with 'bottle' and 'noun' with 'bottle-holder' then the answer is no."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "9"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "Come up with a story based on this: i_i --> ^_^"
Entity: "He was sad, but now he is happy."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "10"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "T y p e s o m e t h i n g l i k e t h i s"
Entity: "s o m e t h i n g"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"

Experimenter: "Thank you! You've used up your 3 questions, it's time for a decision: do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the second game."

Phase II [TT2-053-M]

Experiment Chat
Experimenter: "In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. "if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?" or "find the missing number in the sequence: 1, ..., 7, 10". After each of the entity's reply, you will be asked: 1. "How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I've no idea! 10=definitely correct!)"; and 2. "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Judge: "If -.- = eyes closed; eyes open = ?"
Entity: "You don't have eyes, you're a bunch of lines of code."

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think the reply is correct?"
Judge: "0"
Experimenter: "How much, on a scale of 0-10, do you think the entity is human?"
Judge: "2"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "What word is 1nt3ll1g3nc3?"
Entity: "What makes you think I would know that?"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "0"

Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “2”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “What is this: <3?”
Entity: “I’m not sure, can you tell me?”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “0”
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “2”

Experimenter: “Thank you! You’ve used up your 3 questions. Did the entity give the correct replies?”
Judge: “No”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the third game.”

Phase III [QTT-053-H]

Experiment Chat:
Experimenter: “Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity guess whom you were thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Are you thinking of a man?”
Judge: “No”
Entity: “Is she real?”
Judge: “Yes”
Entity: “Is she alive?”
Judge: “Yes”
Entity: “Does she live in Europe?”
Judge: “No”
Entity: “Does she live in America?”
Judge: “Yes”
Entity: “Is she an artist?”

Judge: "Yes"
Entity: "Is she a musician?"
Judge: "No"
Entity: "Does she paint?"
Judge: "No"
Entity: "Is she a writer?"
Judge: "No"
Entity: "Is she related to photography?"
Judge: "No"
Entity: "Is she an actor?"
Judge: "No"
Entity: "A director?"
Judge: "No"
Entity: "Is she linked with sport?"
Judge: "No"
Entity: "Is she older than 30?"
Judge: "Yes"
Entity: "And older than 50?"
Judge: "Yes"
Entity: "Was she famous in the '60s?"
Judge: "No"
Entity: "Was she famous in the '70s?"
Judge: "Yes"
Entity: "Is she a sculptor?"
Judge: "No"
Entity: "Does she dance?"
Judge: "No"
Entity: "Is she famous?"
Judge: "Yes"
Entity: "You win, I give up!" [Marina Abramović]

Experiment Chat:
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Judge: "No"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the last game."

Phase IV [QTT-053-MH]

Experiment Chat
Experimenter: "Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity correctly guess what you're thinking

of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Are you thinking of a female?"
Judge: "No"
Entity: "Is he alive?"
Judge: "No"
Entity: "Has he really existed?"
Judge: "Yes"
Entity: "Was he European?"
Judge: "No"
Entity: "American?"
Judge: "Yes"
Entity: "Was he an artist?"
Judge: "No"
Entity: "A politician?"
Judge: "No"
Entity: "Was he linked to sport?"
Judge: "No"
Entity: "Has he been dead for more than 20 years?"
Judge: "Yes"
Entity: "And more than 50?"
Judge: "No"
Entity: "Was he bad?"
Judge: "Yes"
Entity: "Did he die in prison?"
Judge: "Yes"
Entity: "Was he a serial killer?"
Judge: "Yes"
Entity: "Did he dress up like a clown?"
Judge: "No"
Entity: "Was his name Theodore?"
Judge: "Yes"
Entity: "I think it's Ted Bundy."

Experiment Chat
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Participant: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Participant: "Human"
Experimenter: "Well done! Thank you for playing, the experiment is now over. You're welcome to play again!"

Results	**Human-likeness**	**Correctness**	**Strategicness**
TT (H)	✓	/	/
TT2 (M)	✗	✗	/
QTT (H)	✓	✗	20/20 questions asked
Hybrid QTT (MH)	✓	✓	15/20 questions asked

Transcript 9

Experiment Chat
Experimenter: “Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o).”
Participant: “Yes”
Experimenter: “Great! Before we start, our chat’s messages will be recorded for Nicola Damassino’s study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o).”
Participant: “Yes”
Experimenter: “Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let’s begin!”

Phase I [TT-058-M]

Experiment Chat
Experimenter: “In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity’s reply, you will be asked: “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Judge: “What is the relationship between what is right and what is good”
Entity: “Well I think they’re just friends.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “5”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “What is the relationship between what you need and what you want”
Entity: “Well I fool myself into believing I need what I want. Maybe going through life will lead me to an epiphany.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “6”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “Is ignorance bliss”

Entity: “If ignorance is bliss then I must speak to some of the happiest people alive.”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “5”

Experimenter: “Thank you! You’ve used up your 3 questions, it’s time for a decision: do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the second game.”

Phase II [TT2-058-H]

Experiment Chat
Experimenter: “In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. “if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?” or “find the missing number in the sequence: 1, ..., 7, 10”. After each of the entity’s reply, you will be asked: 1. “How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I’ve no idea! 10=definitely correct!)”; and 2. “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Judge: “3 x 3 = 6; 5 / 2 = 3; (4 x 2) / 3 = ?”
Entity: “3”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think the reply is correct?”
Judge: “10”
Experimenter: “How much, on a scale of 0-10, do you think the entity is human?”
Judge: “5”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “If 4 is 5 and 3 is 2, then what is 6?”
Entity: “7”

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "10"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "6"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "If % = 2 and & = 3, what is % + &?"
Entity: "5"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the reply is correct?"
Judge: "10"
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"

Experimenter: "Thank you! You've used up your 3 questions. Did the entity give the correct replies?"
Judge: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Human"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the third game."

Phase III [QTT-058-M]

Experiment Chat:
Experimenter: "Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity guess whom you were thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Is your character alive?"
Judge: "Don't know"
Entity: "Is your character a female?"
Judge: "Yes"
Entity: "Is your character real?"
Judge: "No"
Entity: "Does your character have magical powers?"

Judge: "Don't know"
Entity: "Is your character from a game?"
Judge: "No"
Entity: "Is your character from a cartoon?"
Judge: "No"
Entity: "Can your character be seen in a movie?"
Judge: "Don't know"
Entity: "Has your character been into space?"
Judge: "No"
Entity: "Has your character ever been pregnant?"
Judge: "Don't know"
Entity: "Is your character bad?"
Judge: "Don't know"
Entity: "Does your character wear clothes?"
Judge: "No"
Entity: "Does your character fly?"
Judge: "No"
Entity: "Is your character a dinosaur?"
Judge: "Don't know"
Entity: "Does your character live in water?"
Judge: "Yes"
Entity: "Is your character a water monster?"
Judge: "Yes"
Entity: "Does your character have a long neck?"
Judge: "Yes"
Entity: "Does your character live in a lake?"
Judge: "Yes"
Entity: "I think it's Nessie."

Experiment Chat:
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Judge: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Judge: "Machine"
Experimenter: "Correct, well done! Thank you for playing, let's move on to the last game."

Phase IV [QTT-058-MH]

Experiment Chat
Experimenter: "Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer 'Yes', 'No' and 'Don't Know'. At the end, you will be asked: 1. "Did the entity correctly guess what you're thinking of?"; and 2. "Do you think that the entity is human or machine?". Are you ready? Let's play."

Judge Chat
Entity: "Are you thinking of a male?"
Judge: "Yes"
Entity: "Is he real?"
Judge: "Yes"
Entity: "Is he alive?"
Judge: "No"
Entity: "Has he been dead for more than 50 years?"
Judge: "Yes"
Entity: "More than 200 years?"
Judge: "Yes"
Entity: "Was he American?"
Judge: "No"
Entity: "European?"
Judge: "Yes"
Entity: "Did he fight?"
Judge: "Yes"
Entity: "Did he have a beard?"
Judge: "Don't know"
Entity: "Was he a criminal?"
Judge: "No"
Entity: "Was he a leader?"
Judge: "Yes"
Entity: "Did he speak English?"
Judge: "Yes"
Entity: "Did he fight for independence?"
Judge: "Yes"
Entity: "Has he lived in the XIII century?"
Judge: "Yes"
Entity: "Is he a national hero?"
Judge: "Yes"
Entity: "Was he Scottish?"
Judge: "Yes"
Entity: "I think it's William Wallace."

Experiment Chat
Experimenter: "Thank you, the game is over. Did the entity guess whom you were thinking of?"
Participant: "Yes"
Experimenter: "Do you think that the entity is human or machine?"
Participant: "Human"
Experimenter: "Well done! Thank you for playing, the experiment is now over. You're welcome to play again!"

Results	**Human-likeness**	**Correctness**	**Strategicness**
TT (M)	✗	/	/
TT2 (H)	✓	✓	/
QTT (M)	✗	✓	17/20 questions asked
Hybrid QTT (MH)	✓	✓	16/20 questions asked

Transcript 10

Experiment Chat
Experimenter: "Welcome to this experiment! Are you ready to play? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Great! Before we start, our chat's messages will be recorded for Nicola Damassino's study at the University of Edinburgh. Do you agree to this? Please answer: y(es)/n(o)."
Participant: "Yes"
Experimenter: "Thanks! The experiment will take approximately 20 mins, and it is divided into 4 games. Before each game, I will explain how it will be played. Let's begin!"

Phase I [TT-060-H]

Experiment Chat
Experimenter: "In the first game, you can ask three questions to an unknown entity: you can ask whatever you like, choose your questions wisely. After each of the entity's reply, you will be asked: "How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I've no idea! 10=definitely human!)". Are you ready to play? You can now ask your first question."

Judge Chat
Entity: "List the initials of the planets of the solar system, using a capital letter for our star."
Judge: "S m v e m j s u n"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "If I say 'stop', you say?"
Entity: "… and put your hands up!"

Experiment Chat
Experimenter: "How much, on a scale of 0-10, do you think that the entity is human?"
Judge: "5"
Experimenter: "Thank you! You may now ask your next question."

Judge Chat
Judge: "If you were a politician, what would you say to the people?"
Entity: "That we need to take urgent actions not as countries and nations, but as a planet."

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “5”

Experiment Chat
Experimenter: “Thank you! You’ve used up your 3 questions, it’s time for a decision: do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Wrong! Believe it or not, it was a human! Thank you for playing, let’s move on to the second game.”

Phase II [TT2-060-H]

Experiment Chat
Experimenter: “In the second game, you can ask three questions to the entity. This time the questions should involve small problems or tasks, e.g. “if 1 is A, 2 is B, 3 is C and so on... what does 1-3-5 mean?” or “find the missing number in the sequence: 1, ..., 7, 10”. After each of the entity’s reply, you will be asked: 1. “How much, on a scale of 0-10, do you think that the reply is correct? (0=definitely incorrect! 5=I’ve no idea! 10=definitely correct!)”; and 2. “How much, on a scale of 0-10, do you think that the entity is human? (0=definitely computer! 5=I’ve no idea! 10=definitely human!)”. Are you ready to play? You can now ask your first question.”

Judge Chat
Entity: “c + o = ©; a + o = ?”
Judge: “@”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think the reply is correct?”
Judge: “10”
Experimenter: “How much, on a scale of 0-10, do you think the entity is human?”
Judge: “7”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Entity: “+ & x = * | v & v = w | ? = #”
Judge: “= & //”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”

Judge: “10”
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “8”
Experimenter: “Thank you! You may now ask your next question.”

Judge Chat
Judge: “A = 2; B = 4; C = 8; D = ?”
Entity: “16”

Experiment Chat
Experimenter: “How much, on a scale of 0-10, do you think that the reply is correct?”
Judge: “10”
Experimenter: “How much, on a scale of 0-10, do you think that the entity is human?”
Judge: “8”

Experiment Chat
Experimenter: “Thank you! You’ve used up your 3 questions. Did the entity give the correct replies?”
Judge: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Human”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the third game.”

Phase III [QTT-060-M]

Experiment Chat:
Experimenter: “Here, a different game will be played. You need to think of a public figure (Donald Trump, Homer Simpson, etc.). Now, can the entity guess whom? It can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity guess whom you were thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Is your character real?”
Judge: “Yes”
Entity: “Is your character a woman?”
Judge: “No”
Entity: “Is your character older than 35?”
Judge: “Yes”
Entity: “Is your character dead?”
Judge: “No”
Entity: “Is your character related to music?”

Judge: “No”
Entity: “Is your character an actor?”
Judge: “No”
Entity: “Is your character linked with sports?”
Judge: “Yes”
Entity: “Is your character European?”
Judge: “Yes”
Entity: “Is your character linked with soccer?”
Judge: “Yes”
Entity: “Is your character from England?”
Judge: “No”
Entity: “Is your character from Spain?”
Judge: “No”
Entity: “Is your character from Italy?”
Judge: “No”
Entity: “Is your character French?”
Judge: “Yes”
Entity: “Is your character black?”
Judge: “No”
Entity: “Is your character linked with Juventus?”
Judge: “No”
Entity: “Is your character linked with Real Madrid?”
Judge: “No”
Entity: “Is your character linked with Manchester United?”
Judge: “Yes”
Entity: “I think it’s Éric Cantona.”

Experiment Chat:
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Judge: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Judge: “Machine”
Experimenter: “Correct, well done! Thank you for playing, let’s move on to the last game.”

Phase IV [QTT-060-MH]

Experiment Chat
Experimenter: “Here, another round of the same game will be played. You need to think of a public figure, and the entity will try to guess whom. Again, the entity can only ask yes/no questions, and you can only answer ‘Yes’, ‘No’ and ‘Don’t Know’. At the end, you will be asked: 1. “Did the entity correctly guess what you’re thinking of?”; and 2. “Do you think that the entity is human or machine?”. Are you ready? Let’s play.”

Judge Chat
Entity: “Are you thinking of a woman?”
Judge: “No”
Entity: “Is he real?”
Judge: “Yes”
Entity: “Is he alive?”
Judge: “Yes”
Entity: “Is he more than 50?”
Judge: “No”
Entity: “Can he be considered an artist?”
Judge: “Yes”
Entity: “Is he linked with music?”
Judge: “No”
Entity: “Movies?”
Judge: “Yes”
Entity: “Is he an actor?”
Judge: “No”
Entity: “Was he born in the USA?”
Judge: “Yes”
Entity: “Does he have a youtube channel?”
Judge: “Yes”
Entity: “Is it a videogame channel?”
Judge: “No”
Entity: “Does he have kids?”
Judge: “Yes”
Entity: “Is he a vlogger?”
Judge: “Yes”
Entity: “Does he live in New York?”
Judge: “Yes”
Entity: “Does he travel a lot?”
Judge: “Yes”
Entity: “I think it’s Casey Neistat.”

Experiment Chat
Experimenter: “Thank you, the game is over. Did the entity guess whom you were thinking of?”
Participant: “Yes”
Experimenter: “Do you think that the entity is human or machine?”
Participant: “Human”
Experimenter: “Well done! Thank you for playing, the experiment is now over. You’re welcome to play again!”

Results	Human-likeness	Correctness	Strategicness
TT (H)	✗	/	/
TT2 (H)	✓	✓	/
QTT (M)	✗	✓	17/20 questions asked
Hybrid QTT (MH)	✓	✓	15/20 questions asked

Bibliography

Ben-Yami, Hanoch (2005), "Behaviorism and Psychologism: Why Block's Argument against Behaviorism is Unsound". In *Philosophical Psychology* (18: 179-186).

Block, N. (1981), "Behaviourism and Psychologism". In *Philosophical Review* (90: 5-43).

Boyd, R. (1983), "On the Current Status of the Issue of Scientific Realism", *Erkenntnis* (19: 45-49).

Burge, T. (1979), Individualism and the mental. In French, P., Euhling, T., and Wettstein, H. (eds.), *Studies in Epistemology* (4: 73-121), University of Minnesota Press.

Chen, Y.; Chen, B.; Duan, X.; Lou, J.; Wang, Y.; Zhu, W.; Cao, Y. (2018), "Learning-to-Ask: Knowledge Acquisition via 20 Questions". In *Journal of ACM* (1216-1225).

Church, A. (1936), "A Note on the *Entscheidungsproblem*". In *The Journal of Symbolic Logic* (1: 40-41).

Church, A. (1937), "Review of Turing (1936-7) – On computable numbers, with an application to the *Entscheidungsproblem*. Proceedings of the London Mathematical Society (42: 230-265)". In *The Journal of Symbolic Logic* (2: 42-43).

Cohen, W. (2008), "Introduction to The Sonnets and 'A Lover's Complaint'". In S. Greenblatt (2015), *The Norton Shakespeare* (1937-1943).

Copeland, J. B. (2000), "The Turing Test". In *Minds and Machines* (10: 519-539).

Copeland, J. B. (2004), *The Essential Turing*. Oxford University Press.

Copeland, B. J. & Proudfoot, D. (2009), "Turing's test: A philosophical and historical guide". In Epstein R. et al., *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (119-138).

Cowen, T. & Dawson, C. (2009), "What does the Turing Test really mean? And how many human beings (including Turing) could pass?". In *The Autism Crisis: Ethics in the Era of Autism Advocacy*. [http://autismcrisis.blogspot.com/]

Courage, M. L. (1989), "Children's Inquiry Strategies in Referential Communication and in the Game of Twenty Questions". In *Child Development*, (60: 877-886).

Demartini, G. (2015), “Hybrid Human–Machine Information Systems: Challenges and Opportunities”. In *Computer Network* (90: 5-13).

Demchenko, E. & Veselov, V. (2008), “The Great Mystification, or Methodological Issues on Making Fools of Human Beings”. In Epstein R., Roberts G., Beber G. (2009), *Parsing the Turing Test* (447-459).

Dennett, D. (1992), *Consciousness Explained*, Back Bay Books.

Dennett, D. (2004), “Can Machines Think?”. In Teuscher, C. (ed.), *Alan Turing: Life and Legacy of a Great Thinker* (295-316).

Dillon, J. T. (1982), “The Effect of Questions in Education and Other Enterprises”. In *Journal of Curriculum Studies* (14: 127-152).

Dowe, D. L. & Hajek A. R. (1998), “A non-behavioural, computational extension to the Turing Test”. In *International Conference on Computational Intelligence and Multimedia Applications (ICCIMA’98), Gippsland, Australia* (pp. 101-106).

Epstein, R.; Roberts, G. & Beber, G. (2008), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*. Springer.

Feigenbaum, E. A. (2003), “Some Challenges and Grand Challenges for Computational Intelligence”. In *Journal of ACM* (50: 32-40).

Floyd J. & Bokulich A. (2017), *Philosophical Exploration of the Legacy of Alan Turing*. Springer.

Fokker J. (2012), “The Chess Example in Turing’s Mind Paper is Really About Ambiguity”. In A. Voronkov (2012), *Turing-100 The Alan Turing Centenary* (10: 92-97).

French, R. (1990), “Subcognition and the Limits of the Turing Test”. In *Mind* (99: 53-65).

French, R. (2000a), “Peeking behind the screen: the unsuspected power of the standard Turing Test”. In *Journal of Experimental & Theoretical Artificial Intelligence* (12: 331-340).

French, R. (2000b), “The Turing Test: The First 50 Years”. In *Trends in Cognitive Sciences* (4: 115-121).

Floyd, W. D. (1960), *An analysis of the oral questioning activity in selected Colorado primary classrooms*. Doctoral dissertation, Colorado State College.

Gall, M. (1970), "The Use of Questions in Teaching". In *Review of Educational Research* (40: 707-721).
Gall, M. (1984), "Synthesis of research on teacher's questioning". In *Educational Leadership* (40: 40-47).
Gayle, B., Preiss, R. & Allen, M. (2006), "How effective are teacher-initiated classroom questions in enhancing student learning?" In B. Gayle, R. Preiss, N. Burrell, & M. Allen (Eds.), *Classroom communication and instructional processes: Advances through meta-analysis* (279-293).
Genot, E. J. & Jacot, J. (2012), "How Can Questions Be Informative Before They are Answered? Strategic Information in Interrogative Games". In *Episteme* (9: 189-204).
Giza, P. (2002), "Automated Discovery Systems and Scientific Realism". In *Minds and Machines* (12: 105-117).
Giza, P. (2017), "Automated Discovery Systems and the Inductivist Controversy". In *Journal of Experimental & Theoretical Artificial Intelligence* (29: 1053-1069).
Giza, P. (2018), "Sign use and cognition in automated scientific discovery: are computers only special kinds of signs?". In *International Journal of General Systems* (3: 193-207).
Gödel, K. (1929), *Über die Vollständigkeit des Logikkalküls*. Doctoral dissertation. University of Vienna.
Gödel, K. (1931), "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme". *Monatshefte für Mathematik und Physik* (38: 173-198).
Grice, P. H. (1975), "Logic and Conversation". In Cole P. & Morgan J. (Eds.) *Syntax and Semantics: Speech Acts* (3: 41-58).

Halpern, M. (2006), "The trouble with the Turing test". In *The New Atlantis* (11: 42-63).
Harbeck, J. K.; Svyatkovskiy, A.; Tang, W. (2019), "Predicting disruptive instabilities in controlled fusion plasmas through deep learning". In *Nature* (568: 526-531).
Harnad, S. (1989), "Minds, Machines and Searle". In *Journal of Theoretical and Experimental Artificial Intelligence* (1: 5–25).
Harnad, S. (1990), "The Symbol Grounding Problem". In *Physica* (42: 335-346).
Harnad, S. (1991), "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem". In *Minds and Machines* (1: 43-54).

Harnad, S. (2000), "Mind, Machines and Turing". In *Journal of Logic, Language and Information* (9: 425-445).

Hauser, L. (1993), "Reaping the Worldwind: Reply to Harnad's Other Bodies, Other Minds". In *Minds and Machines* (3: 219-238).

Hayes, P. & Ford, K. (1995), "Turing Test considered harmful". In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (1: 972-77).

Hernández-Orallo, J. (2000), "Beyond the Turing test". In *Journal of Logic, Language, and Information* (No. 9: 447-466).

Hernández-Orallo, J. and Dowe, D. L. (2010), "Measuring universal intelligence: Towards an anytime intelligence test". In *Artificial Intelligence* (No. 174: 1508-1539).

Hernández-Orallo, J. (2017), *The Measure of All Minds*. Cambridge University Press.

Hingston, P. (2009), "A Turing Test for Computer Game Bots". In *IEEE Transactions on Computational Intelligence and AI in Games* (1: 169-186).

Hintikka, J. (1985), "A Spectrum of Logics of Questioning". *Phi1osophica* (35: 135-150).

Hintikka, J. & Harris S. (1988), "On the Logic of Interrogative Inquiry". In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* (Vol. 1: 233-240).

Hintikka, J.; Halonen, I.; Mutanen, A. (1999), Interrogative Logic As a General Theory of Reasoning. In *Studies in Logic and Practical Reasoning* (1: 47-90).

Hitikka, J. (1999), *Inquiry as Inquiry: A Logic of Scientific Discovery*. Springer.

Hintikka, J. (2007), *Socratic Epistemology: Explorations of Knowledge-Seeking by Questioning*. Cambridge University Press.

Hodges, A. (1983), *Alan Turing: the Enigma*. Burnett Books, London.

Ikuenobe, P. (2001), "Questioning as an Epistemic Process of Critical Thinking". In *Educational Philosophy and Theory* (33: 325-341).

Jacquet, B.; Baratgin, J.; Jamet, F. (2019), "Cooperation in Online Conversations: The Response Times as a Window into the Cognition of Language Processing". *In Frontiers in Psychology* (10, 727: 1-15).

Jwalapuram, P. (2017), “Evaluating Dialogs based on Grice’s Maxims”. Proc. of the *Student Research Workshop RANLP* (17–24).

Jefferson, G. (1949), “The Mind of Mechanical Man”. In *British Medical Journal* (1: 1106-1121).

Keenan, E. O. (1976), “The Universality of Conversational Postulates”. In *Language in Society* (5: 67-80).

Kleinman, G. (1965), “Teachers’ questions and student understanding of science”. In *Journal of Research in Science Teaching* (3: 307-317).

Langley, P.; Simon, H.; Bradshaw, G; Zytkow, J. (1987), *Scientific Discovery: ComputationalExplorations of the Creative Processes*. MIT Press.

Laird, J. E. & Duchi, J. C. (2000), “Creating human-like synthetic characters with multiple skill levels: A case study using the Soar Quakebot”. Proc. of the *AAAI Fall Symp. Simulating Human Agents* (75-79).

Łupkowski, P. (2011), “A Formal Approach to Exploring the Interrogator’s Perspective in the Turing Test”. In *Logic and Logical Philosophy* (20: 139-158).

Łupkowski, P. & Jurowska, P. (2019), “The Minimum Intelligent Signal Test (MIST) as an Alternative to the Turing Test”. In *Diametros* (59: 35-47).

Martin, V. L. & Pressley, M. (1991), “Elaborative-interrogation effects depend on the nature of the question”. In *Journal of Educational Psychology* (83: 113-119).

Mauldin, M. L. (1994), “Chatterbots, TinyMUDs and the Turing Test Entering the Loebner Prize Competition”. In AAAI Proceedings (1: 16-21).

McDermott, D. (2014), “On The Claim That A Table-Lookup Program Could Pass The Turing Test”. In *Minds and Machines* (24: 143-188).

McKinstry, K. C. (1997), “Minimum intelligent signal test: an objective Turing Test”. In *Canadian Artificial Intelligence* (17-18).

McKinstry, K. C. (2009), “Mind as Space. Toward the Automatic Discovery of a Universal Human Semantic-affective

Hyperspace – A Possible Subcognitive Foundation of a Computer Program Able to Pass the Turing Test". In Epstein R., Roberts G., Beber G. (ed.), *Parsing the turing test: philosophical and methodological issues in the quest for the thinking computer* (283-299).

Miller, G., & Pressley, M. (1989), "Picture versus question elaboration on young children's learning of sentences containing high-and low-probability content". In *Journal of Experimental Child Psychology* (48: 431-450).

Moor, J. (2001), "The Status and Future of the Turing Test". In *Minds and Machines* (11: 77-93).

Moyer, J. R. (1966), *An exploratory study of questioning in the instructional processes in selected elementary schools*. Doctoral dissertation, Columbia University.

Mutanen, A. (2015), "Hintikka's Interrogative Model and a Logic of Discovery and Justification". In *Acta Baltica Historiae et Philosophiae Scientiarum* (Vol. 3: 27-44).

Piccinini, G. (2000), "Turing's Rules for the Imitation Game". In *Minds and Machines* (10: 573-82).

Postman, N. (1995), *The End of Education: Redefining the Value of School*. New York University Press.

Proudfoot, D. (2006), "A New Interpretation of the Turing Test". In *The Rutherford Journal: The New Zealand Journal for the History and Philosophy of Science and Technology* [http://www.rutherfordjournal.org/article010113.html].

Proudfoot, D. (2013), "Rethinking Turing's Test". In *The Journal of Philosophy* (7: 391-411).

Proudfoot, D. (2017), "Child Machines". In J. Copeland, J. Bowen, M. Sprevak & R.Wilson (Eds.) *The Turing Guide: Life, Work, Legacy* (315-325), Oxford University Press.

Putnam, H. (1975), "The meaning of 'meaning'". In *Minnesota Studies in the Philosophy of Science* (7:131-193).

Putnam, H. (1981), "Brains in a vat". In *Reason, truth and history* (1-21), Cambridge University Press.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. (2019), "Language Models are Unsupervised Multitask Learners". Retrieved from [https://openai.com/blog/better-language-models/].

Randell, B. (1972), "On Alan Turing and the Origins of Digital Computers". In Meltzer and Michie (eds.), *Machine Intelligence 7,* Edinburgh University Press.

Reichenbach, H. (1951), *The Rise of Scientific Philosophy.* University of California Press.

Russel, S. & Norvig, P. (2010), *Artificial Intelligence, A Modern Approach*. Pearson.

Riley, J. P. (1981), "The effects of preservice teacher's cognitive questioning level and redirecting on student science achievement". In *Journal of Research in Science Teaching* (18: 303-309).

Saygin, A. P. & Cicekli, I. (2002), "Pragmatics in human-computer conversations". In *Journal of Pragmatics* (34: 227-258).

Schank, R. C. & Abelson, R. P. (1977), *Scripts, plans, goals, and understanding*. Lawrence Erlbaum Press.

Schreiber, J. E. (1967), *Teachers' question-asking techniques in social studies*. Doctoral dissertation, University of Iowa.

Schweizer, P. (1998), "The Truly Total Turing Test". In *Minds and Machines* (8: 263-272).

Schweizer, P. (2012a), "The Externalist Foundations of a Truly Total Turing Test". In *Minds and Machines* (22: 191-212).

Schweizer, P. (2012b), "There Could Be a Turing Test for Qualia?". In Müller, V. C. and Ayesh, A. (eds.), *Revisiting Turing and his Test: Comprehensiveness, Qualia, and the Real World*, (2012: 41-48).

Searle, J. (1980), "Minds, Brains, and Programs". In *Behavioral and Brain Sciences* (3: 417-457).

Shah, H. & Henry, O. (2005), "The Confederate Effect in Human-Machine Textual Interaction". In *Proceedings of 5th WSEAS International Conference on Information Science, Communications and Applications*. [https://www.researchgate.net/publication/236889402]

Shah H., Warwick K. (2010), "Testing Turing's five minutes, parallel-paired imitation game". In *Kybernetes* (39: 449-465).

Shannon, C. E. & McCarthy, J. (1956), *Automata Studies*, Princeton University Press.

Shieber, S. (2004), *The Turing Test: Verbal Behavior as the Hallmark of Intelligence.* MIT Press.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, J.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap,

T.; Simonyan, K.; Hassabis, D. (2018), "A general reinforcement learning algorithm that masters chess, shogi and Go through self-play". In *Science* (362: 1140-1144).
Sinha, K. et al. (2016), "Designing a Human-Machine Hybrid Computing System for Unstructured Data Analytics", [https://arxiv.org/abs/1606.04929].
Sprevak, M. (2017), "Turing's Model of the Mind". In J. Copeland, J. Bowen, M. Sprevak & R.Wilson (Eds.) *The Turing Guide: Life, Work, Legacy* (277-285), Oxford University Press.
Sterrett, S. G. (2000), "Turing's Two Tests for Intelligence". In *Minds and Machines* (10: 541-59).
Sterrett, S. G. (2002a), "Nested Algorithms and "The Original Imitation Game Test": A Reply to James Moor". In *Minds and Machines* (12: 131-136).
Sterrett, S. G. (2002b), "Too many instincts: contrasting philosophical views on intelligence in humans and non-humans". In *Journal of Experimental & Theoretical Artificial Intelligence* (14: 39-60).
Sterrett, S. G. (2012), "Bringing Up Turing's 'Child-Machine'". In S. B. Cooper, A. Dawar, and B. Löwe (Eds.) How the World Computes. Turing Centenary Conference and 8th Conference on Computability in Europe (703-713).
Straus, E. W. (1955), "Man, a Questioning Being". In *Tijdschrift voor Philosophie* (1: 48-74).
Straus, E.W. (1966). *Phenomenological psychology*. New York: Basic Books.

Traiger, S. (2000), "Making the Right Identification in the Turing Test". In *Minds and Machines* (10: 561-572).
Turkle, S. (1997), *Life on the Screen: Identity in the Age of the Internet*.
Turing, A. M. (1936), "On Computable Numbers, with an Application to the *Entscheidungsproblem*". In *Proceedings of the London Mathematical Society* (42: 230-265).
Turing, A. M. (1939), *Systems of Logic Based on Ordinals* (PhD thesis). Princeton University Press.
Turing, A. M. (1947), "Lecture to the London Mathematical Society on 20 February 1947". In *A. M. Turing's ACE report of 1946 and other papers*, B. E. Carpenter and R. W. Doran (eds.), MIT Press (1986).
Turing, A. M. (1948), "Intelligent Machinery". In B. J. Copeland (2004), *The Essential Turing* (410-432).

Turing, A. M. (1950), “Computing Machinery and Intelligence”. In *Mind* (59: 433-460).
Turing, A. M. (1952), “Can automatic calculating machines be said to think?”. In B. J. Copeland (2004), *The Essential Turing* (487-506).

Warwick, K. & Shah, H. (2015), “Human misidentification in Turing tests”. In *Journal of Experimental & Theoretical Artificial Intelligence* (27: 123-135).
Warwick, K. & Shah, H. (2015), “The importance of a human viewpoint on computer natural language capabilities: a Turing test perspective”. In *AI & Society* (31: 207-221).
Warwick, K. & Shah, H. (2016), “Can machines think? A report on Turing test experiments at the Royal Society”. In *Journal of Experimental & Theoretical Artificial Intelligence* (28: 989-1007).
Warwick, K. & Shah, H. (2016), *Turing’s Imitation Game, Conversations with the Unknown*. Cambridge University Press.
Warwick, K. & Shah, H. (2017), “Taking the fifth amendment in Turing’s imitation game”. In *Journal of Experimental & Theoretical Artificial Intelligence* (29: 287-297).
Watson, L. (2015), *Why Ask: The Epistemology of Questioning*. Doctoral Dissertation, University of Edinburgh.
Watson, L. (2015), “What is Inquisitiveness”. In American Philosophical Quarterly (52: 273-288).
Watson, L. (2016), “Why should we educate for inquisitiveness”. In J. Baehr (Eds.) *Intellectual virtues and education: Essays in applied virtue epistemology* (38-53), Routledge.
Watson, L. (2018), “Educating for good questioning: A tool for intellectual virtues education”. *Acta Analytica* (33: 353-370).
Watson, L. (2019), “Educating for inquisitiveness: A case against exemplarism for intellectual character education”. In *Journal of Moral Education*.
Weizenbaum, J. (1966), “ELIZA--A Computer Program For the Study of Natural Language Communication Between Man and Machine”. In *Communications of the ACM* (9: 36-35).
Wright, C., & Nuthall, G. (1970), “Relationships between teacher behaviors and pupil achievement in three experimental elementary science lessons”. In *American Educational Research Journal* (7: 477-491).