UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE FILOSOFIA
PROGRAMA DE PÓS-GRADUAÇÃO EM FILOSOFIA

Bernardo Gonçalves

Machines will think: structure and interpretation of Alan Turing's imitation game

São Paulo

2020

Bernardo Gonçalves

Machines will think: structure and interpretation of Alan Turing's imitation game

Tese apresentada ao Programa de Pós-graduação em Filosofia do Departamento de Filosofia da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

Supervisor: Prof. Dr. Edelcio de Souza

São Paulo
2020

Catalogação na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

Gm

Gonçalves, Bernardo Nunes
Machines will think: structure and interpretation of Alan Turing's imitation game / Bernardo Nunes Gonçalves ; orientador Edelcio G. de Souza. - São Paulo, 2020.
289 f.

Tese (Doutorado)- Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Departamento de Filosofia. Área de concentração: Filosofia.

1. Alan Turing. 2. Can machines think?. 3. The imitation game. 4. Thought experiment. 5. Artificial intelligence. I. de Souza, Edelcio G., orient. II. Título.

Bernardo Gonçalves

Machines will think: structure and interpretation of Alan Turing's imitation game

Tese apresentada ao Programa de Pós-graduação em Filosofia do Departamento de Filosofia da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

Trabalho depositado. São Paulo, 10 de dezembro de 2020:

Prof. Dr. Edelcio de Souza
Universidade de São Paulo
Supervisor

Prof. Dr. Osvaldo Pessoa Jr.
Universidade de São Paulo
Examinador

Prof. Dr. Pío Garcia
Universidad Nacional de Córdoba
Examinador

Prof. Dr. Richard Staley
University of Cambridge
Examinador

São Paulo
2020

To

Prof[a]. Rosane Caruso

(Federal University of Espírito Santo, Computer Science Department),

Prof. José Karam Filho,

Prof. Fábio Porto

(National Laboratory for Scientific Computing, Petrópolis, Brazil),

and Prof. Pablo Mariconda

(University of São Paulo and Scientia Studia),

for their true passion for science.

To

Tânia, Francisco (in memoriam), Maria José, Leonardo, Scheilla, Marildo and Carolina,

for their unconditional love.

Indeed, one may compare the nerves of the machine I am describing with the pipes in the works of these fountains [of a Renaissance garden].

— René Descartes (1633)

The very limited character of the machinery which has been used until recent times (e.g. up to 1940) [...] encouraged the belief that machinery was necessarily limited to extremely straightforward, possibly even to repetitive, jobs.

— Alan M. Turing (1948)

Abstract

GONÇALVES, B. Machines will think: structure and interpretation of Alan Turing's imitation game. 2020, 289 f. Thesis (Doctorate) — Faculdade de Filosofia, Letras e Ciências Humanas. Departmento de Filosofia, Universidade de São Paulo, São Paulo, 2020.

Can machines think? I present a study of Alan Turing's iconic imitation game or test and its central question. Seventy years of commentary has been produced about Turing's 1950 proposal. The now legendary "Turing test" has grown a life of its own in the tradition of analytic philosophy with at best loose ties to the historical imitation tests (1948-1952) posed by Turing. I shall examine the historical and epistemological roots of Turing's various versions of imitation game or test and make the case that they came out from within a dialogue, in fact a scientific controversy, most notably with physicist and computer pioneer Douglas Hartree, chemist and philosopher Michael Polanyi, and neurosurgeon Geoffrey Jefferson. Placing Turing's views in their historical, social and cultural context, I shall reclaim their scientific and philosophical value for the sake of the discussion in the years to come. My study is organized according to three main philosophical problems whose analyses are backed by a subsidiary chronology of the concept of machine intelligence in Turing's thought (1936-1952).

The first problem I will address is the identification of Turing's specific ambition which led him to announce that machines will think. War hero and brilliant mathematician, he challenged the conventional wisdom of what machines really were or could be and prophesized a future pervaded by intelligent machines which may be seen as a dystopia just as much as a utopia. I shall examine Turing's profile and take special interest in the way he was seen by his contenders.

In the second problem, over and above the mere proposal of a test for machine intelligence, I will study Turing's proposition "machines can think" and its implied existential hypothesis — "there exists (will exist) a thinking machine" — from a point of view of the history of the philosophy of science. Unlike traditional readings of Turing, I found that Turing held a non-obvious realist attitude towards the existence of a mechanical mindbrain which he conjectured to frame the human and whose digital replica he intended to build in the machine.

Turing's 1950 paper has been acknowledged as a complex and multi-layered text. Opposing views can be identified in the literature relative to the question on whether or not Turing proposed his imitation test as an experiment to decide for machine intelligence. I shall call this the Turing test dilemma and address it as my third and main problem. My findings suggest that Turing cannot have proposed his imitation game as something other than a thought experiment. And yet its critical and heuristic functions within the mind-machine controversy are striking.

Keywords: Alan Turing. Can machines think?. The imitation game. Thought experiment. Artificial intelligence.

Resumo

GONÇALVES, B. As máquinas vão pensar: estrutura e interpretação do jogo da imitação de Alan Turing. 2020, 289 pp. Tese (Doutorado) — Faculdade de Filosofia, Letras e Ciências Humanas. Departmento de Filosofia, Universidade de São Paulo, São Paulo, 2020.

Podem as máquinas pensar? Apresento um estudo do icônico jogo ou teste da imitação de Alan Turing e sua questão central. Setenta anos de comentários foram produzidos sobre a proposta de Turing de 1950. O já lendário "teste de Turing" ganhou vida própria na tradição da filosofia analítica, na melhor das hipóteses com uma fraca relação com os testes da imitação históricos (1948-1952) apresentados por Turing. Examinarei as raízes históricas e epistemológicas das várias versões do jogo ou teste da imitação de Turing e argumentarei que elas surgiram no interior de um diálogo, de fato uma controvérsia científica, notavelmente com o físico e pioneiro da computação Douglas Hartree, o químico e filósofo Michael Polanyi e o neurocirurgião Geoffrey Jefferson. Situando as ideias de Turing em seu contexto histórico, social e cultural, reivindicarei seu valor científico e filosófico para benefício da discussão nos anos que virão. Meu estudo está organizado em três problemas filosóficos principais cujas análises são apoiadas por uma cronologia subsidiária do conceito de inteligência de máquina no pensamento de Turing (1936-1952).

O primeiro problema que abordarei é a identificação da ambição específica de Turing que o levou a anunciar que as máquinas vão pensar. Herói de guerra e matemático brilhante, ele desafiou a sabedoria convencional sobre o que as máquinas realmente eram ou poderiam ser, e profetizou um futuro permeado por máquinas inteligentes que pode ser visto tanto como uma distopia quanto como uma utopia. Examinarei o perfil de Turing, com especial interesse pela maneira como ele foi visto por alguns de seus antagonistas. No segundo problema, para além da mera proposta de um teste para inteligência de máquina, estudarei a proposição de Turing "as máquinas podem pensar" e sua hipótese existencial implicada — "existe (existirá) uma máquina pensante" — do ponto de vista da história da filosofia de ciência. Diferentemente das leituras mais tradicionais de Turing, meu achado é de que Turing mantinha uma atitude realista não óbvia em relação à existência de uma mente-cérebro mecânica que ele conjecturou moldar o humano e cuja réplica digital ele pretendia construir na máquina. O artigo de Turing de 1950 foi reconhecido como um texto complexo e multifacetado. Visões opostas podem ser identificadas na literatura quanto à questão de se Turing propôs ou não seu teste da imitação como um experimento para decidir pela inteligência da máquina. Chamarei esse problema (terceiro e central que abordo neste estudo) de "o dilema do teste de Turing." Meus achados sugerem que Turing não pode ter proposto seu jogo da imitação como outra coisa senão um experimento mental. E no entanto suas funções crítica e heurística, no interior da controvérsia mente-máquina, são marcantes.

Palavras-chave: Alan Turing. Podem as máquinas pensar?. Jogo da imitação. Experimento mental. Inteligência artificial.

List of Figures

List of Tables

Contents

Introduction

Suppose there is some mechanism — let us call it *think*$_2$ — that could be built into a machine and make it talk freely and significantly better than parrots. The mechanism could be different in essence than the faculty that is embodied in us humans — call it *think*$_1$ —, yet it might be just enough, and for the sake of argument let us consider that it is enough, to make the machine indistinguishable from humans in terms of talking, that is, to make think$_2$ indistinguishable from think$_1$ when it comes to speech. In the context of a key public debate that took place in England (1949-1952), British mathematician Alan Turing (1912-1954) suggested that no one knew what think$_1$ really is nor how different would it be from conjectured think$_2$, so it would *not* be nonsense at all to include think$_2$ into the extension of what we simply call *thinking*. In fact, in doing so, Turing committed to a much earlier proposal of French philosopher René Descartes's which implied unrestricted conversation as a proxy for thinking — notwithstanding Descartes presumed the existence of a true talking machine not to be practically possible. For Turing, the actual existence of a thinking machine was rather an empirical claim. It was a matter of allowing for some tractable advances in the science and technology of modern computing to take place. In effect, Turing posed a bold existential hypothesis that had, and still has, a fairly clear empirical content. It is this hypothesis that justifies his claim for an extension in use of the term "thinking." And it is this hypothesis that I invite the reader to consider as the object of this dissertation.

In his seventeenth-century views, Descartes was following the principles of his mechanistic science or natural philosophy. In the ending of Part V of his *Discourse on the method* published in 1637 (I shall consider Cottingham et al.'s 1985 [1637] edition), Descartes held that it was not practically possible for either machines or animals to have as much diversity of behavior, verbal or otherwise, as we humans do. For him, their mechanisms were limited to one function each. To put it differently, machines were necessarily special-purpose. But to explain the contrasting human capability to behave reasonably in all possible situations, Descartes attached to his mechanistic science the metaphysical category of the rational soul, which is endowed with (general-purpose) reason and granted to each of us humans by God. For Cottingham, "Descartes's adherence to the thesis of the incorporeality of the mind" was motivated by a "triad of considerations, theological, metaphysical and scientific," and "it would be difficult or impossible to single out any one as having the primacy in structuring his own personal convictions" (1992, p. 252). And yet, in regard to the rational soul, Cottingham pondered (p. 253), it is only at the end of his exposition in the *Discourse* that Descartes "tacked on it." "The soul," Cottingham argued, is "invoked to account for the phenomena of thought and language that appeared to Descartes, for empirical reasons, radically resistant to mechanistic explanation." In any case, that triad seems to have helped Descartes to escape persecution. That he feared persecution we know from his November 1633, February and April 1634 letters to Mersenne. In those letters Descartes directly

addressed Galileo's condemnation in 1633 by the Congregation of the Holy Office. In the (1991 [1633]) letter, he resonated with Mersenne that Galileo's "views about the movement of the earth were condemned as heretical," and continued "I would not wish, for anything in the world, to maintain [my views] against the authority of the Church." He then concluded "I desire to live in peace and to continue the life I have begun under the motto 'to live well you must live unseen'."

In his twentieth-century views, Turing also relied on a new science. The science of digital computers, however, led him towards the opposite direction, for these machines hold the crucial property of *universality* — they can imitate any special-purpose machine. Any amount of diversity one may like to see in the behavior of a general-purpose machine would be a matter of granting it sufficient storage capacity (Turing also called it "memory") and, most importantly, a suitable program. In fact, digital computers can have their capabilities extended by the addition of all sorts of special-purpose programs or tables of instructions. For instance, a specific form of digital computer that Turing called in (1936) a "universal computing machine" is a general-purpose machine that can be given in its tape a program to make it break coded messages, another program to play chess well, and so on. The universal machine was defined so as to imitate under strict "discipline" each special-purpose machine that is encoded into its tape as input. But even after a first universal machine was announced in *Nature* in September (1948) to have been actually built and set operational, Turing never claimed that it could succeed at the human-level conversation task — for the universal Turing machine as defined in 1936 lacked "initiative." In order not to be limited by the situations its designer could imagine *a priori*, the machine would have to partially trade discipline for initiative. As described in Turing's (1948-1952) outlines of his research project, the machine would have to be made to learn for itself — like the mindbrain of a human child —, with no need to be turned off and reset before it could react to a newly posed challenge. So thought Turing, and he did not fear persecution. Turing consciously challenged the conventional wisdom of the time as caught in common phrases that he cited such as "acting like a machine," or "purely mechanical behaviour," or "you cannot make a machine to think for you." He also opposed the admonitions of neurosurgeon Geoffrey Jefferson (1886-1961), who back in (1949a) referred to the state of the art in neuroscience and also to Descartes in seek of support for his views, and even urged that "the concept of thinking like machines lends itself to certain political dogmas inimical to man's happiness [and] erodes religious beliefs that have been mainstays of social conduct" (p. 1107). Turing conjectured and argued that it was possible to program a machine so that it would think. From 1948 to 1952, he discussed his views by means of various forms of an experiment that he named in his famous (1950) paper the *imitation game* and also referred to as "my test."

After all, can machines think? This dissertation addresses Alan Turing's imitation game or test and its central question in some of its core scientific, philosophical, historical and social aspects. There is already an enormous literature on the so-called Turing test. So why to propose yet another interpretation of it? I will offer an answer that considers four different limitations and gaps that I identify in the secondary literature. I shall then complete my answer with a general

justification of my dissertation in view of the current state of affairs relative to the science and technology of artificial intelligence and its current and future impact in society and culture. I will then proceed to introduce the dissertation itself in terms of its chapter structure and the key elements that compose my studies.

First, Turing scholarship so far, it can be observed, has been developed mostly within the so-called analytic tradition of philosophy. Most of the secondary literature on Turing pays little attention to history. The now legendary "Turing test" seems even to have grown a life of its own with at best loose ties to the historical imitation tests proposed by Turing. This may also be due to the irregular availability of sources over time. Some of the key Turing sources have not been available or widely visible to the research community until more recently. For instance, Turing's core 1948 piece *Intelligent machinery* has only been declassified by the British National State and published by then members of the National Physical Laboratory Evans and Robertson in (1968). Three new primary Turing sources have only been published by Jack Copeland in (1999). Minute notes of an event that I propose was crucial in the evolution of Turing's concept of machine intelligence has only appeared when his contemporary Wolfe Mays eventually decided to publish it in (2000). Even having these latest sources available for over twenty years now, nevertheless, analyses still fall short in taking into account the historical, social, and cultural context of Turing's in the late 1940's and the early 1950's in Cold War Britain. A crucial development in this connection is the mind-machine controversy that took place in England (1949-1952), which has barely been noticed in connection with the famous Turing test. For instance, let us consider its seventeen year-old dedicated entry in the *Stanford Encyclopedia of Philosophy* (OPPY; DOWE, 2020 [2003]), and which just received revision a few months ago. Only now in its August 2020 revision it started to acknowledge some relationship between Descartes and Turing's views on how to distinguish men or humans from machines and other animals. Geoffrey Jefferson, Douglas Hartree and Michael Polanyi's names are barely mentioned. And these, as I shall present in due course, are three crucial players in the scientific controversy where Turing's imitation game was born. Their contentions are the origin of nearly all of the nine objections that Turing formulated and rebutted in his 1950 paper. Indeed, the secondary literature barely takes notice of a historical fact that is sheerly manifest in the Turing sources of those years, namely, that Turing's most famous and frequently quoted passages have been delivered by him *from within a dialogue*. Turing was not a prolific writer and he did not leave us prolegomena. However, thanks to the foundational work of a few Turing scholars in collecting primary and secondary sources, we have available a historical record of events that influenced Turing's discourse and yet they have long been neglected in the secondary literature.

Second — and not an unrelated factor —, the secondary literature on Turing's imitation game has been growing at the crossroads of science, history, philosophy, sociology, anthropology and fiction that for seventy years now. However, the interpretation of Turing's 1950 proposal is still remarkably controversial. In fact, Turing's 1950 paper has been acknowledged as a complex and multi-layered text. Some interpreters have seen in Turing's test a decisive experiment and

sometimes even the holy grail of artificial intelligence research. And yet, even among those that saw in it the proposal of an actual — definite and practical — scientific experiment, further heterogeneity arises, for there are supporters *and* critics of the significance of Turing's test as such. Most supporters in this class either dismissed or shrank the element of gender imitation in the test. Most critics in the same class contended that Turing proposed an operational definition of intelligence and a form of behaviorism that is reductive of the human mind. Others yet have seen in Turing's proposal not an actual experiment at all but either just "a joke" or at most a historical manifesto for artificial intelligence with no scientific content inside. Some interpreters in this latter class acknowledged gender imitation as an important element in the structure of the test but considered that it rather testifies against its seriousness. Overall, two opposite positions can be identified relative to the question on whether or not Turing wanted to propose and/or did in fact propose his test as an actual experiment to decide for machine intelligence. This exegetical quagmire still lives. I shall refer to and address it in this dissertation as the Turing test dilemma.

Third, underlying Turing's proposal of his test one should in principle expect to be able to locate his actual views and the nature of his research program. As known, in the 1950 text Turing avoided to directly address his original question, whether machines can think. But in later sources he did address it quite explicitly. And yet there seem to be no clearly available answers to the questions: did Turing actually mean some concept of "thinking" or "intelligence" over and above his verbal behavior test? If so, what did Turing mean exactly, say, if we distinguish epistemological and ontological stances? What was Turing's view about the human mind? What specific theses or hypotheses did he posit, if any? In this dissertation I shall directly address these questions on the basis of an extensive discussion of key passages from Turing sources all the way from the outset of his concept of machine intelligence to its endgame (1936-1952).

Fourth, Turing has been widely recognized as a brilliant mathematician, natural scientist and philosopher. He was appointed a Fellow of the Royal Society and an Officer of the Order of the British Empire for his wartime services and is in fact widely viewed today as a war hero. But he also challenged conventional wisdom of what machines really were or could be, and even suggested releasing them from their duties as slaves. For that he has been accused of Promethean irreverence and been associated, among other things, by critics ranging from his biographer Andrew Hodges (2012 [1983], p. 521) to computer scientists such as Patrick Hayes and Kenneth Ford (1995, p. 976), with the image of Dr. Frankenstein. From some of his contemporary interlocutors he received antagonism also at a personal level, even if but subtly, for the non-conformist stripe of his views in Cold War Britain. A reader of Victorian novelist Samuel Butler, he prophesized a future world pervaded by intelligent machines which may be seen as a dystopia just as much as a utopia. In his (1950) paper, for instance, he "hope[d]" that "machines will eventually compete with men in all purely intellectual fields" (p. 460). In the presence of such diverse facts, one may ask: what intellectual profile did Turing have? What ambitions did him have which led him to announce that machines will think? These are also questions that I will address in this dissertation.

In general, I shall also prompt that Turing's thoughts, words and deeds had large and widespread effects in the formation of our age (AGAR, 2001; COPELAND, 2012). Born out of his pure mathematics in (1936), Turing's ideas rapidly changed warfare, the factory world of industry and labor, and society altogether. His ideas can be argued to deserve attention from fresh and diverse perspectives indeed as new developments continue to take place in this twenty-first century. In particular, recent advances in the science and technology of intelligent machines (today known as the field called artificial intelligence, or AI for short) brought back a debate about the future of such machines in society and culture. The current studies on AI and society nevertheless, pay little if any attention at all to the public debate that took place in the United Kingdom and the USA in the turn from the 1940's to the 1950's. Besides Norbert Wiener and other cyberneticians, Turing was a leading figure and pioneer back then. And Turing, as always in his life, took his own original approach to it. He considered the logical possibilities and limits of (future) machine capabilities, and made foundational contributions to the debate about the future of AI that, I shall hold, are yet to be fully appreciated. A reconstruction of Turing's argument about the future of AI in society and culture must build upon his computer science and should also connect elements of history, philosophy and even literature. While that would for sure require a larger project, I hope to be able in this dissertation to make a first self-contained step in that direction.

Daniel Dennett offered in (2006 [1984]) an interpretation of Turing's test. He drew attention to what is in my view a central point about Turing's famous question which had not yet received enough consideration back then (and perhaps only very recently has started to receive):

> Can machines think? This has been a conundrum for philosophers for years, but in their fascination with the pure conceptual issues they have for the most part overlooked the real social importance of the answer. It is of more than academic importance that we learn to think clearly about the actual cognitive powers of computers, for they are now being introduced into a variety of sensitive social roles, where their powers will be put to the ultimate test: in a wide variety of areas, we are on the verge of making ourselves dependent upon their cognitive powers. (DENNETT, 2006 [1984], p. 295)

It is also in that spirit that this dissertation addresses Turing's test. My specific goal is to contribute with a new interpretation of Turing's views that pays attention to their social relevance, and above all, an interpretation that is distinctively historiographical. My general (and more distant) goal is to contribute to the bulk of knowledge assembled so far by Turing scholars towards positioning Alan Turing beyond the field of analytic philosophy and the mind-body problem, beyond the history of mathematics, computing and cognitive sciences, then effectively as a contemporary philosopher-scientist in dialogue with a central issue for modern philosophy — the status of machines as the other face for the status of humans. Descartes's beast-machine thesis has been one of the foundations of modern philosophy and science, and an excellent assembly of dualism. Essentially, that was the thesis that humans are ontologically different from machines and other animals because the former, but not the latter, are able to think. For Descartes, reason was the

unique belonging of a rational soul. And the presence of thinking could be tested by trying to establish with the inquired entity a reasonable and unrestricted conversation. In his historical imitation tests (1948-1952), indeed as we shall see, Turing met precisely the terms required by Descartes. He thus challenged all of these cartesian theses, and in this century we may know whether he will have succeeded.

Throughout this dissertation I shall introduce and make use of a few philosophical distinctions in view of clarifying what is at stake in the scientific and philosophical dispute about thinking machines. There is a first distinction that I wish to prepare the reader to keep in mind from the start. It was quite early in the public debate about thinking machines that Max Newman — a former teacher of Turing at the University of Cambridge, and longstanding collaborator and witness of Turing's work — offered this differentiation about the imitation of *kinds* of thought by machines, which I suggest us to call, for short henceforth, *Newman's distinction*:

> There is evidently a danger here that extravagant powers will be credited to these devices [electric automatic computing machines], and conclusions drawn too rapidly about biological analogues; but some caution is also necessary in coming to final conclusions in the opposite direction, on the basis of our knowledge of the behaviour of a small pilot model of this very new kind of machine [...]. The first question that will have to be asked is not "Can *all* kinds of thought, logical, poetical, reflective, be imitated by machines?" but "Can *anything* that can be called 'thought' be so imitated and, if so, how much?" (NEWMAN, 1949, no emphasis added)

Newman formulated with mathematical elegance the problem of what intellectual capabilities we are supposed to attribute to the machines. He distinguished two variants of the problem in terms of universal ("all") and existential ("any") predication. His distinction provides in my view a well-reasoned scheme for the study of Turing's views and their reception. Mostly, it has not been used to provide structure for discussion in the reception of Turing's views in general, and of Turing's (1950) paper in particular. It shall, however, accompany my studies in this dissertation.

The dissertation, entitled *Machines will think: structure and interpretation of Alan Turing's imitation game*, is organized in two parts according to Table 1. It is mostly in discussion with the specialized literature on the Turing test and the philosophy of science, and to a lesser extent with the literature on the philosophy of mind. But readers familiar with any one of these topics or fields, I hope, shall find it readable. Also to support a more general readership, I have included an appendix chapter (§A) that gives a systematic chronology of the development of the concept of machine intelligence (1936-1952) in Turing's thought. Facts and events are weaved in narrative for a reconstruction of Turing's reasoning from within its social and historical context. The chapter is optional, intended mostly for readers less familiar with Turing's works and for readers that seek historiographical rigor. It can be profitable, for example, as a detour to gain more depth into Turing's historical and social context after going through the first chapter (§1) and before the second and third chapters (§§2, 3). In the main part (the thesis), in any case, I will often refer to sections of the appendix chapter to source details from their historical and social

Table 1 – Structure of this dissertation.

Main Part	**Machines will think**
Chapter 1	Alan M. Turing (1912-1954): prophet of the machines
Chapter 2	Turing's hypothesis on the existence of thinking machines
Chapter 3	Turing's test is a thought experiment in science
Appendix	**Chronology of Turing's ideas**
Chapter A	Machine intelligence in Turing's thought (1936-1952)

context. Also, my corpus of primary Turing sources and some of the secondary ones can be found by skimming through the table of contents of the appendix (§A). The reader may observe that the primary Turing sources on machine intelligence have different statuses. Only a few, notably his seminal (1936) and (1950) papers, are texts that he consciously wrote and submitted to peer-review for publication. Others are either internal reports written and delivered to industrial and military institutions belonging to his National State, or popular oral communications that he delivered to a professional society (London Mathematical Society) and to the general public (through BBC radio's public service). Also in light of my mid- and long-term goal of presenting Turing as a contemporary philosopher-scientist, it seems remarkable to me to note that some of the most important of his expositions were nothing of the sort of cultivated essays or treatises but just popular radio broadcasts. Turing, despite being a relatively shy person and a particularly specialized intellectual, seems to have enjoyed talking plaining but sharply, directly to the general public. It is interesting to note, for instance, something that his mother Mrs. Sara Turing once related (2012 [1959], p. 100). Turing seemed to feel shy about joining her in listening to his own performance in the May 1951 BBC radio broadcast. For me this is informative that he cared about how he must have sounded in his announcements about the future of the humankind.

As known, lying in the forefront of Turing's famous 1950 paper is his proposal of the imitation game or test for machine intelligence. Turing's test is, in effect, my central problem and will be addressed in the third chapter (§3). The first and the second chapters (§§1, 2) are where I will develop two related aspects of my main argument on Turing's 1950 proposal. Related to my specific approach to interpret Turing's test is my goal of drawing attention to what is perhaps less visible in Turing's 1950 paper, namely, a bold conjecture that Turing posited on the existence of thinking machines and his views about their future in society and culture. Turing's 1950 proposal is often construed in the related literature as a philosophical *thesis*, so to speak, on mind and machine. And it is true that Turing submitted and published his paper in a journal for the philosophy of mind. This mostly philosophical perspective has produced a voluminous literature, and yet Turing's test and its related view are, as mentioned, still very controversial. What if, though, we look at Turing's 1950 proposal from a philosophy of science point of view? I mean it in the sense of the analytic tradition as well but more with, say, the colors of the Vienna Circle of understanding philosophy specifically as a second-order discipline upon the empirical sciences. That is, my studies of Turing's views in this dissertation shall suggest that he may be

better understood if not seen as if he was trying to come up with any new doctrines but as if reflecting upon the new science of computing he had himself created. This is precisely my take here, for I have found in Turing's views, along with his famous test, a scientific *hypothesis*.

In the first chapter (§1), I will study Turing as a thinker. By now, several biographies have appeared on Turing, which the reader may refer to for an acquaintance with his life. From sort of a micro-biography point of view, as it seems to me nonetheless, some questions are yet in the open about the intellectual profile of this thinker who assumed a leading role in a public controversy that took place in England (1949-1952) on whether machines can think. I shall take special interest in the way Turing was seen by some of his contenders. Turing posited a bold hypothesis on mind and machine and held views of a non-conformist stripe in the context of Cold War Britain and, in particular, the period that became known as the Second Red Scare (1947-1957). It was before this background that Turing has been attacked also at a personal level, even if but subtly. His perhaps most authoritative biographer Andrew Hodges accepted and endorsed, curiously enough as we shall see, some very personal views of Jefferson's about Turing, in spite that Jefferson was perhaps Turing's main antagonist. I have thus found to be desirable to try to clear the ground first, by examining key passages from Turing sources and by recovering the social and cultural context of the mind-machine controversy, in preparation to digging deeper into his scientific, philosophical and social views relative to machine intelligence. I will address the problem of identifying the specific (Promethean) ambition that led him to announce that machines can think, and will think. Turing thought, I will try to show, that digital computers were scientific instruments, just like Galileo's telescopes. They would do great service to psychology in the study of the human brain. But he also drew attention to the advent of superintelligent machines, or the possibility of machines to outstrip our intellectual powers and take control. This was, for him, not a certain event but indeed a true possibility. Now, if he thought so and still wanted to build intelligent machines, does that mean that Turing collaborated towards a dystopian future? I think not. I shall argue that he was rather raising a precautionary voice. In the presence of the crass chauvinism of some of his contenders, however, he could not but establish epistemological relations between the possibility of *intelligent* machines and the refutation of what he saw as species and gender biases of his contenders. For this reason, I think, Turing's prophesized future pervaded by intelligent machines may be seen as a dystopia just as much as a utopia. It depends on the social and cultural background of who is seeing. *Superintelligent* machines were for him a true but distant possibility that pushed his argument to the limit from both epistemological and social stances. Turing also expressed a dislike towards the dismissal of the ontological distinction between the natural and the artificial. The evidence I have collected is unfavorable to associating Turing with, say, modern transhumanist movements.

In the second chapter (§2) I will study Turing's proposition "machines can think" and its implied existential hypothesis on thinking machines from a point of view of the philosophy of science. I shall examine it through the lens of a clear-cut distinction between Turing's epistemology and his ontology of thinking or intelligence. By these categories I mean, as usual,

a theory of knowledge and a theory of being. So, concerning Turing's hypothesis, if we look at it from an epistemological point of view, we are trying to understand Turing's proposed view on how to justify the attribution of thinking or intelligence to an entity as a knowledge claim. If we look at it in turn from an ontological point of view, we are trying to understand Turing's proposed view on what thinking or intelligence (really) is. I am not aware of any previous explicit reference to such a distinction relative to Turing's test and hypothesis. My hope is that it will be fruitful towards an understanding of Turing's views and their reception in the secondary literature. I shall also look into Turing's philosophical attitude towards his hypothesis. This amounts to exploring his statements, their tone, his expectations, their targets and so on. My identification of Turing's hypothesis as an existential scientific hypothesis led me to look at the history of the philosophy of science and, in particular, the classical text of Herbert Feigl (1950). The latter turned out to be a most fruitful framework to study both Turing's hypothesis itself and its interpretations in the secondary literature. Contrary to common readings of Turing, and specially those that construe him as a behaviorist, I found that Turing held a realist (and physicalist) attitude towards the existence of a mechanical mindbrain which he conjectured to frame the human and whose digital replica he intended to build in the machine.

In the third chapter (§3) I will study the central problem of interpreting Turing's famous imitation game or test. I shall also try to reconcile opposing views of it from the secondary literature towards resolving I called above the Turing test dilemma. Turing explicitly referred to "my test" both in his 1950 paper and in later sources. But more in the forefront of his narrative in 1950 was his "imitation game." In the related science and philosophy secondary literature one may come across references to both "[the] Turing[']s test" and "Turing's imitation game," and the former turns out to be much more common (for instance, to the point of naming the related entry in the *Stanford Encyclopedia of Philosophy* that has been cited above). Focus on the latter but not on the former is usually a clear sign that the writer takes Turing's 1950 proposal as less significant from a scientific and philosophical point of view. Now, as the reader may have noticed from the title of this dissertation, my option is to emphasize Turing's imitation game. This is not because I will reproduce the pattern just mentioned, quite the opposite. As known, Turing was an irreverent thinker. And I will rather shed light on the scientific and philosophical significance of his irreverent game. My study of the imitation game (or test) as a thought experiment shall unveil just this, for the structure and functions of the imitation game as such address non-obvious and quite important scientific and philosophical issues indeed. Along these lines and by building upon existing interpretations of the Turing test I aim at contributing a new interpretation that holds promise to make sense of opposing views of it in the secondary literature and, at best, I hope, reclaim its scientific and philosophical value. For an example, the controversial gender question that appears in the imitation game, I have found, plays a key function which is to encode in wit a rebuttal to a serious argument posed by Jefferson in his (1949a) Lister Oration about the physiology of behavior. Now, the presentation of an imaginary scenario to opponents within a scientific controversy in order to persuade them towards conceptual change is a distinctive mark

of thought experiments. This is just what Turing did, and the structure of the imitation game as such addresses other assumptions of Jefferson's and of others,' most notably physicist and computer pioneer Douglas Hartree (1897-1958) and chemist and philosopher Michael Polanyi (1913-1976). In fact, it is in this connection comes a central result of my studies ever since I got started three years ago. Remarkably enough, except for some side notes in passing that refer to the imitation game as a "thought experiment," as far as I know there is not a single such study in the secondary literature. It was thus my initial goal to explore a construction of Turing's imitation as a thought experiment and to assess its fruitfulness. I ended up though finding a result that is arguably stronger than that. It turned out to me that (i) a rigorous exegesis of Turing's 1950 text, together with (ii) knowledge of a series of related historical events and (iii) knowledge of an elementary epistemological concept which we know Turing was aware of, altogether, is suggestive that Turing's imitation game cannot have been proposed by him as something other than a thought experiment, no matter he was conscious about that or not. Thought experiments in science may be feasible to be run, and this seems to be the case of Turing's. And yet, one may ask: is running Turing's imitation game a sensible scientific project? In the second chapter of this dissertation we shall examine it in depth, together with all the elements that compose my study of Turing's imitation game as a thought experiment. I shall draw on Ernst Mach's analysis of the nature of thought experiments in science, and also on some historical aspects of Galileo's famous falling-bodies thought experiment which, I hold, can inform the current situation concerning Turing's.

Finally, I will present an overview of the dissertation and the key results in a Conclusion.

Part I

Machines will think

1 Alan M. Turing (1912-1954): prophet of the machines

Alan Turing (1912-1954), according to a key contemporary figure in a condolence letter to Mrs. Sara Turing after her son's death, was "a sort of scientific Shelley" (2012 [1959], p. 58).

Turing was born British and elected Fellow of the Royal Society (FRS) in March 1951 under sponsorship of Bertrand Russell and Max Newman, who wrote Turing's obituary (1955). The first Turing biographies were his mother Mrs. Sara Turing's (2012 [1959]) and a most comprehensive one by Andrew Hodges (2012 [1983]). Several other biographies have appeared since the 2000's (LEAVITT, 2006; COPELAND, 2012; TURING, 2015; SWINTON, 2019), which the reader may refer to for an acquaintance with Turing's life. Alan, the child viewed as eccentric in the British public education system, would graduate in mathematics at Cambridge University in 1934; be elected fellow of King's College, Cambridge, early in 1935 at only his 22; go to America and earn a doctorate degree in mathematical logic from Princeton University in (June) 1938; recruit himself to the British Foreign Office and help the Allies win the Second World War from September 1939 on; join the National Physical Laboratory to build Britain's "electronic brain" in 1945; and eventually, in postwar Manchester, assume role of revolutionary thinker who challenged (and whose remaining views still challenge) conventional wisdom.

My intention here will be to present an intellectual-biography sketch or mini-biography of Turing with focus on the role he assumed in the public controversy that took place in England (1949-1952) on whether machines can think. This period comprised almost three of the last five years of Turing's life. I shall take special interest in the way he was seen by some of his contenders. The image of Turing as a scientific Shelley was given by Sir Geoffrey Jefferson (1886-1961), "master of the neurosciences and man of letters" (SCHURR, 1997). Jefferson is important because he stood out, most notably at the time, against Turing's views on machine intelligence. And Turing bothered to give him replies, such that Jefferson can be said to have been Turing's primary philosophical opponent. In 1949, when they first met, Jefferson (63) was near double Turing's age (37). Scientific Shelley is a startling portrait of Turing that Jefferson rendered. I think that it has significance for a sketch of Turing's profile, as discussed next.

1.1 Problem and chapter structure

Biographer Andrew Hodges takes for granted (2012 [1983], p. 439) that it must have been Percy Shelley the one that Jefferson associated Turing to. This is, I think, most likely correct. Yet Jefferson's reference may have had a dual sense and been related as well to Mary Shelley (as known, Mary and Percy Shelley were married), for reasons that I elaborate in what follows. In

fact, I found that Hodges himself bought and endorsed Jefferson's delicate double suggestion right away in his interpretation of Turing. He presumed: "[t]here was a Shelley in him, but there was also a Frankenstein — the proud irresponsibility of pure science, concentrated in a single person" (p. 521). I think this is no minor issue, and it deserves serious examination.

I shall refer to Jefferson's ambiguous association of Turing with (the) Shelley(s) as *the problem of identifying Turing's specific (Promethean) ambition*. I am concerned with having as most an accurate understanding of Jefferson's view of Turing as possible and will keep up pursuing it. And yet the question has a value of its own independently of Jefferson's perspective, and this is for the sake of a more rigorous and accurate biographical portrait of Turing.

Jefferson had made admonitions and other dubious suggestions in relation to Turing. For instance, when Turing was elected FRS in March 1951 (at his 38), Jefferson, who had been elected FRS just four years before (at his 61), sent him a letter:

> I am so glad; and I sincerely trust that all your valves are glowing with satisfaction, and signalling messages that seem to you to mean pleasure and pride! (but don't be deceived!). (TURING, 2012 [1959], p. 101)

Playing with words, Jefferson suggests Turing to be a machine. He also alluded to some kind of behaviorism, perhaps Gilbert Ryle's logical behaviorism, as if it were related to Turing's views, precisely the views that they, Jefferson and Turing, had in dispute since June 1949. So, in the occasion of congratulating Turing for a unique recognition by the scientific community of their country, Jefferson delivered a joke packed with their contentious issues inside.

Percy Shelley once left in the register of Chamonix's Hôtel de Londres this threefold autobiographical note: "I am a lover of mankind, a democrat and an atheist."[1] Atheism aside, would Jefferson have thought of Turing as a lover of mankind and a democrat? Percy Shelley, as known, was one of the major English romantic poets together with his friend Lord Byron and others. Shelley is said not to have been as influential in his own time (1792-1822) as he was in the three or four generations next. Considered a radically progressive thinker, most publishers and journals in his time declined to publish his work, if for nothing else, for fear of being arrested for blasphemy or subversion. From 1818 to 1822 (until his death) Shelley wrote his masterpiece *Prometheus unbound* (SHELLEY, 1959 [1818-1822]), a four-act lyrical drama in reference to Greek dramatist Aeschylus' trilogy conventionally called the *Prometheia*. The three plays were *Prometheus Bound*, *Prometheus Unbound* and *Prometheus the Fire-Bringer*. The second and the third plays were preserved only in fragments. Essentially, the *Prometheia* is concerned with the torments of the Greek mythological figure Prometheus, who defies the gods and gives fire to humanity, for which he is subjected to suffering and eternal punishment by Zeus (Jupiter). Unlike in Aeschylus' version, Percy Shelley exhibited a Prometheus that, with the help of his lover Oceanid Asia, manages to get released from the oppression of Zeus.

[1] Cf. "Shelley's shocking 'atheist' declaration rediscovered after 200 years". Available at <http://www.trin.cam.ac.uk/news/shelleys-shocking-atheist-declaration-rediscovered-after-200-years/>. Access on 13 mar. 2020.

Roughly at the same time, as known, Mary Shelley (1797-1851), in some non-obvious collaboration with her husband (for one aspect of it, they both wrote prefaces to each other's work), wrote the famous *Frankenstein: the modern Prometheus* (2012 [1818]). In her fiction, Dr. Victor Frankenstein infuses the spark of life in an otherwise inert body only to get later horrified by his creature. Dr. Frankenstein's attitude, as represented by Mary Shelley, is a reference to Promethean disobedience of a different sort. Blind in his pursuit of modern science, Dr. Frankenstein loses control of his creature and its actions. Now, understanding the connections between the two Shelley's pieces is a problem of its own which received, for instance, an informative analysis by critical biographer Richard Holmes (2011). Although Percy's Prometheus is a modern figure, it is actually Mary's the one that engaged in an outright profane attitude towards the creation myth and offered a dominating image about the artificial creation of life. A fairly simplified view of the differences between the figures of Prometheus in *Prometheus unbound* and *Frankenstein* will suffice to my purpose here, which is as follows.

The fire stolen by Percy Shelley's (enlightened and progressive) Prometheus is a symbol for the emancipation of humans by the knowledge and the abilities that will enable them to shape their own future, now independent of the gods. This is different from Dr. Frankenstein, or the (unwary and outrageous) modern Prometheus of Mary Shelley's, who just in the course of his reckless research gave fire inadvertently to a creature whose (in)human status is a problem. From 1816 to 1818, some critics note, Mary Shelley seems to have started to develop some deep concerns about modern science. She may even have read, in that period, the first official version of Goethe's Faust Part I, published in 1808. The apparently unlimited possibilities open by science would, for Mary Shelley, be uncanny. For the sake of this rough image of contrast, Mary's Prometheus is dangerous and may be approximated with Goethe's Faust.

So the question is posed: in the end, what kind of Prometheus did Jefferson suggest to associate with Turing? As mentioned, I thematize this as the problem of identifying Turing's specific Promethean ambition. I will revisit and address it in the central moment of this chapter (§1.5). In what follows (§1.2) I shall introduce Turing's own view and what he himself called "Promethean irreverence," and also have a closer look at Jefferson's view of him. Also informative will be to acquire a perspective on Turing's use of irony (§1.3) and his style of reasoning (§1.4). I shall arrive at my proposed interpretation of Turing's ambition (§1.5) and imply a specific image of him. Then I will proceed to complete my exposition of Turing's intellectual profile and role as prophet of the machines by presenting his plea (§1.6) and his fate (§1.7). I will outline an analytical summary collecting my findings and the key points developed (§1.8), and conclude with an epilogue (§1.9) and chapter acknowledgements (§1.10).

1.2 Turing's irreverence

Mentions of "Frankenstein" appeared in the English press in 1946 in connection with the advent of modern computing machines, when a buzz came out about scientists who were building an "electronic brain." We learn from Hodges (2012 [1983], p. 347) that this term seems to have been first used publicly on 31 October 1946 by British statesman Louis Mountbatten (1946). In fact, Turing was hired by the National Physical Laboratory (or NPL for short) in 1945 after the Second World War just to build "a brain" (cf. §A.3.2). And Mountbatten would have been briefed by the NPL staff in 1946, we learn from the historical account given by Turing scholars Diane Proudfoot and Jack Copeland in (2018). For these authors, "Turing's views were probably the inspiration for much of Mountbatten's [1946] address" (p. 27).

Turing did refer to Prometheus explicitly. It was in his (2004 [1948]) discussion of objection "(b)" to the possibility of machine thinking (p. 410). Had him given it a name, it could well have been "the Promethean objection." Turing formulated it this way: "[a] religious belief that any attempt to construct such [intelligent] machines is a sort of Promethean irreverence." He then responded: "being purely emotional, [it does] not really need to be refuted" (p. 411). And yet in the same text (2004 [1948]), in a section he entitled "man as machine," Turing also wrote:

> One way of setting about our task of building a 'thinking machine' would be to take a man as a whole and try to replace all the parts of him by machinery. He would include television cameras, microphones, loudspeakers, wheels and 'handling servo-mechanisms' as well as some sort of 'electronic brain'. (TURING, 2004 [1948], p. 420)

Later, in his (1950) paper, Turing enlarged and reshaped his 1948 list of objections. Then even though not as quite directly, in his discussion of the closely related objection now named "the theological objection," Turing addressed with some irony the same charge:

> In attempting to construct such [intelligent] machines we should not be irreverently usurping His [God's] power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates. (TURING, 1950, p. 443)

Jefferson may not have had access to Turing's 1948 report but it is very unlikely that he did not read Turing's 1950 text. (He is addressed directly and indirectly several times there, and they have been in touch in the period, at least up to January 1952.) In any case, it may have been from a more total perspective or Gestalt that Turing reminded Jefferson of the Shelleys developed out of their encounters. Unfortunately the records of their meetings and mutual impressions are relatively limited. We can grasp from Mrs. Sara Turing's biography that in addition to a few public meetings, there may have been a few private and informal meetings too. We know for sure of one, recollected by Jefferson himself in the same sympathy letter to Mrs. Sara Turing which is quoted by Lyn Irvine (Newman's wife) in her foreword to the biography:

> He was a wonderful chap in many ways. I remember how he came to my house late one evening to talk to Professor J.Z. Young and me after we had been to a meeting in the Philosophy Department here, arranged by Professor Emmet. I was worried about him because he had come hungry through the rain on his cycle with nothing but an inadequate cape and no hat. After midnight he went off to ride home some five miles or so through the same winter's rain. He thought so little of the physical discomfort that he did not seem to apprehend in the least degree why we felt concerned about him, and refused all help. It was as if he lived in a different and (I add diffidently, my impression) slightly inhuman world. (IRVINE, 2012 [1959], p. xx)

So, in Jefferson's view, as dramatic as it may look, "it was as if [Turing] lived in a different" and "slightly inhuman world." I think this is intriguing. And one shall not dismiss Jefferson's view of Turing as idiosyncratic of a personal impression or born out of some obvious rivalry. To add another source who engaged with Turing in the discussion on whether machines can think, let us consider Wolfe Mays (1912-2005). He participated with Turing in an October 1949 seminar held in the Department of Philosophy in the first edition of the same meeting mentioned by Jefferson and happens to have saved with him minute notes of the meeting whose author is unknown. This seminar is a critical event in Turing's intellectual life, as we shall see later (§3.5). For a dedicated account of the seminar in itself, the reader may refer to (§A.4.2). One year later, as Turing's 1950 paper was about to appear, according to Mays himself he was asked by Gilbert Ryle to write Turing a reply (2001). And so he did. In his text, which would be rejected by Ryle and only get published in (1952), Mays alluded to "[t]he paradoxical Frankenstein nature of the machine-mind" (p. 150) and implied to Turing the designation of a "mechanical necromancer" (p. 153).

Inhumane views of Turing's "inhuman world" may, of course, be contrasted with some of the warm testimonies of Turing's friends and work colleagues, which can be found in plenty in the Turing biographies cited above. For example, one may consider stories such as this one from Robin Gandy, who received doctoral supervision in mathematical logics from Turing at the University of Cambridge and was one of Turing's best friends:

> When we were engaged on war work, I always thought him a bit austere but at Cambridge I was enchanted to find how human he could be, discussing mutual friends, arranging a dinner-party, being a little vain of his clothes and his appearance. One of my happiest memories is of him and Nick Furbank and me playing a complicated game of hide-and-seek in the Botanical Gardens by moonlight. (TURING, 2012 [1959], p. 119)

Gandy's note about Turing being more austere when engaged on war work, on the one hand, and being dazzling to the extent of playing hide-and-seek in the gardens by moonlight in peace time, on the other, should not pass to the reader unnoticed. In any case, I will concentrate on views such as Jefferson and Mays' because they are germane to the reception of Turing's views on machine thinking. If contemporary fellows saw Turing that way — as a "mechanical necromancer" or "as

if he lived in a different" and "slightly inhuman world"—, then I take to be an important question to ask: why so? What is in such view? I shall keep up exploring this throughout this dissertation.

Turing thought that machines can think. And he had a project to build an intelligent machine. Now, regarding the general position of contenders such as Jefferson, Mays and others, I would like to distinguish these two interpretation options (and my intent is less to reduce their views and more to foster clarity in my own discussion):

(i) they took *Turing's views and project to be feasible* and have dangerous ethical and social implications; so Turing was, for them, sort of a mad scientist in the white coat that could create some monster in the laboratory and then had to be stopped;

(ii) they rather took them to be unfeasible and significantly mixed with some sort of wishful thinking; so Turing was, for them, sort of a deceiver whose actual damage could not be the release of an artificial creature but rather to mislead the public opinion.

I would like to promptly suggest that Jefferson, Mays and other contenders were ambivalent and slippery in between the two positions above. Hodges reported that "Alan would refer to Jefferson as an 'old bumbler' because he never grasped the machine model of the mind" (2012 [1983], p. 439; no primary or secondary source is cited, neither is informed when Turing would have said that.) If Hodges is right, then, for Turing, the dispute had a basis on Jefferson's confusions about the logic of computing. In any case, let us fixate attention on interpretations (i) and (ii) above.

In relation to interpretation (i), did Turing consider ethical and social implications of building and deploying in the society thinking machines? For a short answer, I say yes — and that is why he exposed himself and spoke out. But in fact, he seems to have been half pleased about the implications of a machine revolution. As Diane Proudfoot nicely put in a rare moment of acknowledging the seriousness of Turing's social comments (2015), "[h]e seemed almost to welcome the possibility of this humiliating lesson for the human race." (I shall return to hers interpretation and those of other Turing scholars soon.) Let us see. In discussing the fifth objection to machine intelligence, "arguments from various disabilities" (*viz.*, various things contenders brought forth as if machines could never do), Turing uttered this twofold delicacy:

> The works and customs of mankind do not seem to be very suitable material to which to apply scientific induction. A very large part of space-time must be investigated, if reliable results are to be obtained. Otherwise we may (as most English children do) decide that everybody speaks English, and that it is silly to learn French. (TURING, 1950, p. 448)
>
> What is important about this disability [being unable to enjoy strawberries and cream] is that it contributes to some of the other disabilities, *e.g.* to the difficulty of the same kind of friendliness occurring between man and machine as between white man and white man, or between black man and black man. (TURING, 1950, p. 448, no emphasis added)

With his peculiar touch of irony (§1.3), Turing addressed two key forms of chauvinism. But Turing considered that it would take several decades, maybe one century, for the impact of intelligent machines to come in society and culture. So, for him, I interpret, his generation would have enough time to study the logical limits of (future) machine capabilities beforehand. That is, Turing's attitude in raising such a strong voice about the social importance of considering the future of intelligent machines — when only a few were paying attention — can also be seen as *precautionary*. I will support and further develop this answer later (§1.5).

Now, moving on to possible interpretation (ii) above, is it likely that Turing was at least in part lost in some sort of mystical dream or desire to pursue an absurd project and deceive public opinion? I think not, as discussed next (§1.3, §1.4, and §1.5).

1.3 Turing's irony

Jack Copeland chose "humour" as one of three words (and the one appearing first) to sum up Alan Turing (2012, p. 1). Turing's remarkable sense of humor can also be noted in the contributions of several of his friends, some of which vividly appearing in Christopher Sykes' BBC documentary (1992). I invite the reader that is less familiar with Turing's biography to appreciate what was, indeed, one of the main facets of his. My specific goal here, though, is to discourage a specific reading of Turing's sense of humor. Both Turing's contemporaries and later commentators took his wit or liked to suggest it as reason to more or less dismiss the seriousness of his views. For instance, back in the highs of the polemic about intelligent machines in England (§A.4.1), along the same lines we have just seen, Turing delivered a non-obvious reply to some strong claims purporting a myriad of things that machines would never be able to do. He then received from the June (1949) editorial of the *British Medical Journal* (henceforth also BMJ) this rejoinder:

> Mr. A. W. [*sic*] Turing, who is one of the mathematicians in charge of the Manchester "mechanical brain," said in an interview with *The Times* (June 11) that he did not exclude the possibility that a machine might produce a sonnet, though it might require another machine to appreciate it. Probably he did not mean this to be taken too seriously [...]. (BMJ, 1949, p. 1129)

I observe that this was the first of a series of *ad hominem* attacks that Turing would receive. As known, these are a form of logical fallacy when the interlocutor rather than the argument is put under scrutiny. Jefferson, as the reader may have guessed at this point, was a master of dressing that as compliment. For instance, he wrote to Turing's mother in the same passage of the sympathy letter previously mentioned: "He was so unversed in worldly ways, so childlike it seems to me, so unconventional, so non-conformist to the general pattern [...] so very absentminded," and completed "[h]is genius flared because he had never quite grown up" (TURING, 2012 [1959]), p. 58). So Turing, according to Jefferson, would have been like that, nothing but innocent.

Against this, I shall argue that Turing knew very well that his views on intelligent machines were bold and unorthodox. But in face of the sometimes unhelpful reaction received, he just could not help himself either in avoiding being fun and provocative. Also, it is worthwhile highlighting that Turing received plenty of provocations of dubious taste. For instance, in the October 1949 philosophy seminar already mentioned, at some point Jefferson is reported to have said "but this is an argument against the machine: do human beings do this kind of thing?"; to what Turing would have replied "yes – mathematicians;" and then a murmur would have followed "are mathematicians human beings?" (Cf. TURING et al., 2005 [1949]). In spite of all that, to the best of my knowledge, Turing has never pushed back along the same lines, that is, by departing from the topic of discussion to engage in *argumentum ad hominem*. It has been the motivation of my research to take Turing's views seriously. For, in spite of a dazzling, hardly offensive touch of humour, he seems to have been as truthful and as clear as possible about his views, whose chronology (§A) suggests that were grounded on his new science of computing.

A notorious example of Turing's integrity, in my view, is his spontaneous initiative of outlining his beliefs in the (1950) paper (p. 442). In that passage Turing also offered two potentially testable predictions about the future (one scientific, the other social and cultural), which I shall discuss later (§3.3). In relation to whether one should take Turing's sense of humor as compromising about the meaning of what he said, I would like to quote two specific remarks. The first one is by Mrs. Sara Turing. After having collected an enormous set of testimonies from Turing's peers and friends in several letters after his death, she summed up:

> There is a marked unanimity of opinion in the letters to me about Alan. Coupled with great admiration for "his profound originality and insight" there is repeated emphasis on his simplicity and integrity and complete "lack of pretentiousness and pomposity." (TURING, 2012 [1959], p. 118)

A related, perhaps more complete note about Turing's profile and manners was given by Turing's fellow mathematician and contemporary at King's College, Denis Williams:

> In intellectual, as in other matters, it was essential to him that everything should ring true. [...] it seems to me precisely this complete intellectual integrity, which, combined with his other gifts, made it reasonable to expect that he would produce results of fundamental importance in his own field. Alan had a delightful sense of humour. He enjoyed elaborating fantastic projects, such as a scheme for faking prehistoric cave paintings, in mock-serious detail, or bringing an over-serious discussion down to earth with a quick colloquial turn of phrase. With him jest and earnestness were often closely intermingled. (TURING, 2012 [1959], p. 91)

Now, if it is fair enough to assume that Turing's sense of humor did not interfere with his intellectual integrity, the question that may yet remain is how should we best understand some of his acutely ironic statements? Meeting with great intellectual but also with — in his own words — "emotional" opposition, Turing made extensive use of irony. Sensing that he was not

properly listened to, some of his communications seem to have been assembled to shock. I think that Turing's irony deserves to be studied and understood, and I leave a more in-depth study as a topic of future work. I also think, however, that we can share some initial understanding of Turing's irony for the benefit of the present discussion.

In fact, Turing thought that machines could think, and suspected that we humans may be, in a certain non-trivial sense (§2.5), like machines. Now, given this, I find in John Price's study of David Hume's irony a remarkable clue towards Turing's:

> When a man was under the intellectual and cultural pressure which Hume experienced he could not respond easily by denunciations, by shouting, or by threats. As a civilized man, Hume would not have responded that way under any circumstance. His method of dealing with those who would persecute him or ostracize him simply because of his religious or philosophical or moral opinions was subtle and effective. Irony gave him a method of operating in a world that found his ideas both strange and shocking: strange because most people were simply unable to handle them, shocking because his scepticism dared to attack the citadel of religion. New ways of thinking about man's place in nature, especially if they do not reassure one's blind faith, are often difficult [...] to tolerate. Irony could at least create artificial tolerance. (PRICE, 1965, p. 4-5)

Clearly, Turing has not been the first bold (British) thinker to make use of first-class irony. And Turing's "new ways of thinking about man's place in nature" were "often difficult to tolerate" indeed. I take Price's interpretation of Hume to be very enlightening when applied to Turing. I claim that Turing's words should neither be understood always literally nor dismissed as plain mockery. Turing's irony was rather a clever form of communication to an intellectual and social environment that could barely listen to him. It came most often as satire, that is, let us say, as irony with a point. Turing applied irony more or less subtly, I interpret, to imitate people's language in ways to expose their stupidity or vices of thought about a subject matter.

It is possible to go even one step further. I know of no use of sense of humor as parody by Turing, that is, let us say, imitation for the sole purpose of a comic effect. Now, does this leave us with no methodological principle to follow in the interpretation of Turing's statements and prone to psychologism? I do not think so. Although Turing relied on irony several times, he was far off from being a hermetic speaker and writer. There are two fairly simple (and hopefully effective) principles that I will try to abide by strictly whenever we are to interpret an ironic statement of Turing's. First, as a form of structuralism, the statement shall be understood on the basis of related (less ironic or non-ironic) passages from Turing sources, and consistently with the whole system of his positions (§A). Second, as a science-studies maneuver, we shall look at Turing's historical and social context and, most specially, whom he was speaking to. In particular, we shall consider the public controversy in which Turing was involved, as mentioned, from 1949 to 1952 (§A.4). I claim that the interpretation of Turing's irony is a tractable problem.

1.4 Turing's style of reasoning

In October 1949, as mentioned, we find Turing engaged in a debate in the philosophy department of his university. We also know that Turing's (1950) paper published one year later would have a profound impact in analytical philosophy. So the question I would like to pose now is: what kind of philosopher was Turing? Mathematician, natural scientist and philosopher, there can be little doubt that Alan Turing was all in one, and that may suffice for a short answer. But such a general view adds little to help us understand Turing and his ideas in depth. My goal here is to study Turing's specific profile from the point of view of the history of science and of philosophy.

Let us start with Max Newman's Royal Society memoir on Turing. I find two astonishing facts in his view of Turing's intellectual biography. First, Newman wrote:

> The varied titles of Turing's published work disguise its unity of purpose. The central problem with which he started, and to which he constantly returned, is the extent and the limitations of mechanistic explanations of nature. (NEWMAN, 1955, p. 256)

One can observe there a remarkable similarity between Turing's natural-philosophy program and René Descartes's. (One may consider, say, Descartes' *The world* and *Treatise on man*, and his *Essays*). Indeed, both Turing and Descartes dedicated themselves to find just *the extent and the limitations* of mechanistic explanations of nature. However, as Newman's follow-up sentence reveals, their approach to address that problem differed significantly. Thus put Newman:

> [Turing's] way of tackling the problem was not by philosophical discussion of general principles, but by mathematical proof of certain limited results: in the first instance the impossibility of the too sanguine [Hilbert's] programme for the complete mechanization of mathematics, and in his final work, the possibility of, at any rate, a partial explanation of the phenomena of organic growth by the 'blind' operation of chemical laws. (NEWMAN, 1955, p. 256)

Descartes's way of tackling the problem, it turns out, was foremostly "by philosophical discussion of general principles," or at least surely not by mathematical proof of limited results. Newman's formulation is, for me, a most elegant take on Turing's intellectual profile as a philosopher.

And yet we can push it forward. In (1994), Ian Hacking wrote a well-known paper about styles of reasoning in science. Hacking retrieved and summarized from historian of science A. C. Crombie's three-volume *Styles of scientific thinking in the European tradition* a list of six styles of thinking, which he preferred to rephrase to "styles of [scientific] *reasoning*" (p. 33, emphasis added). According to Hacking, Crombie's styles result from dense studies and are worked out from the history of Western science "in painstaking detail" (p. 34). Since Hacking's paper, the study of styles of reasoning became an active topic of research in the history and philosophy of science. For my purpose here, however, the Crombie-Hacking styles will suffice. They are particularly interesting to refer to in relation to Turing because, as Hacking puts it, "Crombie

includes mathematics among the sciences, which is where they belong." (Yet Hacking wanted to emphasize that "styles do not determine a content, a specific science;" since "there is only a very modest correlation between [the listed styles] and a possible list of fields of knowledge," p. 34). The six Crombie-Hacking styles are:

> (a) The simple method of postulation exemplified by the Greek mathematical sciences.
> (b) The deployment of experiment both to control postulation and to explore by observation and measurement.
> (c) Hypothetical construction of analogical models.
> (d) Ordering of variety by comparison and taxonomy.
> (e) Statistical analysis of regularities of populations, and the calculus of probabilities.
> (f) The historical derivation of genetic development. (HACKING, 1994, p. 34)

I would like to show that (at least) the first three styles have been well practiced by Turing. First, I shall say that Turing obviously mastered style (a), both as result of his early practice during his education (cf. the Turing sources gathered by Hodges in 2012 [1983], Chapter 1), and later as a result of the professional mathematical training he received at the University of Cambridge. Turing proved plenty of mathematical theorems, having started early from his teens. Additionally, Turing also exercised style (a) in the domain of his empirical studies of the mind as we shall see in this dissertation. In his (1950) paper, he even wrote:

> The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any unproved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research. (TURING, 1950, p. 442)

Turing discussed it generally. However, as he was writing that passage, he was dealing with a specific conjecture, namely, the real-world existence of a thinking machine (§2). Turing exercised the method of postulation in both domains, analytical and empirical, indeed. While, on the one hand, Turing's practice of style (a) in mathematics is sort of a trivial vindication; on the other hand, it is quite interesting to observe how that practice, once turned habit or custom, became key in his empirical studies about mechanistic explanations of the mind. It has let Turing to acknowledge and distinguish with ease whether he was dealing with "proved facts" or "conjectures," as appeared in the above passage. And this leads us to his practice of style (b).

Turing was a mathematician, yet an empiricist. His trust in experiment as the true source of knowledge about nature is thus manifested in this key passage of his (1950) paper:

> The reader will have anticipated that I have no very convincing arguments of a positive nature to support my views. [...] The only really satisfactory support that can be given for the view expressed at the beginning of §6 [a prediction about a machine being able to play well a simplified form of the imitation game], will be that provided by waiting for the end of the century and then

> doing the experiment described. But what can we say in the meantime? What steps should be taken now if the experiment is to be successful?
> (TURING, 1950, p. 454-5)

Turing then proceeded to describe a research agenda to address the problem of how to program a machine so that it could play the imitation game well. Now if Turing was an empiricist who proposed the imitation game and saw value in it as an experiment, it remains open though the question, *what kind of experimenter was Turing?* To answer that question, we shall better break apart the Crombie-Hacking formulation of (b), which covers the deployment of experiment *both* (i) "to control postulation," and (ii) "to explore by observation and measurement."

Turing did find value in the setup of experiments "to explore by observation." He relied often on *initial or preliminary experiments*. For instance, Mrs. Sara Turing quoted *The Manchester Guardian*'s 11 June 1954 obituary, which collected these words:

> Turing took a particular delight in problems, large or small, that enabled him to combine mathematical theory with experiments he could carry out, in whole or part, with his own hands. He was ready to tackle anything which combined these two interests. (TURING, 2012 [1959], p. 118)

(A paradigmatic example of this, we shall see in §3.4, is Turing's 1948 description of his initial experiment on making a machine to play chess.) From preliminary experiments, Turing leaped to non-obvious abstractions. These would be subject then, as suggested by Newman in the quotation above, to mathematical formulation and proof of certain limited results. To my knowledge, Turing did not rely on experiments "to explore by measurement," at least not — say, in the way of his fellow Royal Society predecessors such as Robert Hooke, Robert Boyle and Michael Faraday, namely, to induce regularities or as a means to establish an empirical fact. Now, in regard to the deployment of experiment "to control postulation," Turing knew that the available technology at the time would not enable him to run the imitation game with any decisive role as he suggested in the quotation above. And yet he did allude to experiment as a way to control postulation. In what sense? Although his most famous experiment — his imitation game or test — is a feasible experiment, I will try to show later (§3) that he did not really conceive it to be a crucial experiment towards confirming or refuting a hypothesis. Rather, Turing mastered the art of running experiments of the mind, in the way of, say, Galileo Galilei, Isaac Newton and Albert Einstein. Interestingly, a related notion has been acknowledged early by Norbert Wiener, who said in (1965 [1948]): "Turing, who is perhaps first among those who have studied the logical possibilities of the machine as *an intellectual experiment*" (p. 13, emphasis added). Turing was, in effect, a master of thought experiment. I will present later his imitation game or test as a thought experiment that has a well-designed structure and performs a dual function (§3).

Finally, let us consider Turing's exhibition of style (c), namely, the "[h]ypothetical construction of analogical models." Turing made extensive use of explanatory analogies and the hypothetical construction of analogical models. A most notable example is the famous abstract

machine that Alonzo Church once called and so eternized the "Turing machine" (1937). That was a (1936) hypothetical construction out of an analogy with a human clerk working with pencil and paper and following strict rules of operation. So, *by observing* the workings of the human mind in a specific (mathematical) activity, Turing constructed a machine model of it. Later he *conjectured* that such machine model might perhaps be extended to explain other human mind behaviors, perhaps even the whole of the "real mind" (1950, p. 454). Turing outlined these possibilities by means of a "skin of an onion" explanatory analogy, as we shall see later (§2.5). Another key example is the imitation game itself, which was conceived out of an analogy with a parlor game that was aired in Britain by BBC in the turn from the 1940's to the 1950's (§3.6).

Altogether, one may observe that Turing's practice of styles (a), (b) and (c), was intermingled. Now, I would like to come back to Newman's formulation of Turing's intellectual profile as we have just seen. I think that it provides suggestive evidence that Turing's views on thinking machines, likable or not, were essentially born out of rigorous scientific inquiry. I do not intend at all to suggest a view of science as value-free, or a view that a scientist such as Turing did not receive broader influences. Rather, my interpretation goes as follows.

Incidentally, Turing posed a serious question to the philosophy of mind. The reaction to it has been, and still up till these days — a strong one. Turing's 1950 proposal, as we shall see later (§3.1), is still very controversial. What I want to emphasize here, however, is that philosophers of the mind, on the one hand, in general, propose and defend (philosophical) *theses*; and Turing, on the other hand, articulated a (scientific) *hypothesis*. What I have been trying to show in this section is that *Turing, as a philosopher, thought like a scientist*.

1.5 Turing's ambition

Machines shall not be condemned to be slaves forever. Thus implied Turing, seemingly for the first time publicly in a 1947 lecture to the London Mathematical Society when he said:

> It is [...] true that the intention in constructing [electronic computing] machines in the first instance is *to treat them as slaves*, giving them only jobs which have been thought out in detail [...] Up till the present machines have only been used in this way. But is it necessary that they should always be used in such a manner? (TURING, 2004 [1947], p. 392-3, emphasis added)

Turing tried to make the point that seeing machines as slaves may be quite a short-minded perspective. Being "slaves" such a strong word — and the master-slave dichotomy was indeed implied overall in Turing's lecture, with multiple occurrences of both words —, one may then ask, *what future did Turing actually want for the machines and how does it relate to mankind?* While the question may seem at first to engage in a form of psychologism, the reader may recall that it actually came from Jefferson's historical testimony the suggestion that there was in Turing an ambition. (§1.1). Following on, we have seen a related discussion by Turing himself (§1.2),

and also have become acquainted with Turing's irony (§1.3) and his style of reasoning (§1.4). It is now time to revisit and address the problem of identifying Turing's Promethean ambition.

The prophet and the ostriches

In 1948 Turing formulated an objection "(a)" to the possibility of machine thinking:

> An unwillingness to admit the possibility that mankind can have any rivals in intellectual power. This occurs as much amongst intellectual people as amongst others: they have more to lose. Those who admit the possibility all agree that its realization would be very disagreeable. The same situation arises in connection with the possibility of our being superseded by some other animal species. This is almost as disagreeable and its theoretical possibility is indisputable. (TURING, 2004 [1948], p. 410)

Turing thus posed a *fact-value demarcation* that is very important to consider for identifying his ambition. In the presence of this, it seems that any criticism of his ethics may have to take a stand and cope with the question of science and values more generally as well. In any case, note that here we are less interested in whether his demarcation would be effective or not (and I am suggesting nothing in this regard), and more interested in interpreting how Turing himself saw it in the turn from the 1940's to the 1950's. He distinguished the problem of studying the possibility of us humans being superseded by machines or other animal species in intellectual power, on the one hand, and whether or not it would be agreeable to someone or to mankind, on the other. Moreover, as a second point that I take from the passage, he said "[t]hose who admit the possibility" — and this count must obviously include himself in — "all agree that its realization would be very disagreeable." I take this to suggest that the advent of (super)intelligent machines to supersede us humans in intellectual power is, for Turing, at least in part, not welcome or agreeable. In any case, for him, what is at stake does not depend on how likable it is. A third takeaway is, for me, that Turing challenged the anthropocentric attitude towards machines and animals that has been established among us at least since René Descartes's beast machine thesis.

Later, in (1950), Turing would write "I believe [...] that no useful purpose is served by concealing these beliefs [on the possibility of intelligent machines]" (p. 442). And he would rephrase his 1948 formulation quoted above and name it the "heads in the sand" objection to machine intelligence (p. 444). Three aspects stand out in my view: first, Turing's insistence on challenging such an objection; second, the 1950 denomination he assigns to it; and third, the fact that Turing knew, of course, that the 1950 text, unlike the 1948 one, would go to the public domain. Altogether, I think that it gives enough to establish Turing's (1950) paper as the moment when he clearly assumes his role as *prophet of the machines*. The reformulated objection ran:

> (2) The 'Heads in the Sand' Objection. "The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so." [...] We like to believe that Man is in some subtle way superior to the rest of creation. It is best if he can be shown to be *necessarily* superior, for then there is no danger

> of him losing his commanding position.
> (TURING, 1950, p. 444, no emphasis added)

Turing spoke out, seemingly in spite of noticing that only a few were open to listen, and he did not seem to fear persecution. Further on in the same text, he concluded:

> We may hope that machines will eventually compete with men in all purely intellectual fields. [...] We can only see a short distance ahead, but we can see plenty there that needs to be done. (TURING, 1950, p. 460)

Let us now recall Newman's distinction from Introduction, on whether we shall attribute to machines the capability to imitate "*all* kinds of thought, logical, poetical, reflective," on the one hand, or "*anything* that can be called 'thought'," on the other hand (1949). In the passage above, Turing referred to "all purely intellectual fields." (He referred to "purely intellectual" as opposed to, say, fields that require situated reasoning about present-state physical circunstances.) So in 1950 Turing had already departed from the restricted variant of the problem in the sense of Newman. Moreover, it is such 1950 use of "hope" followed by a reference to "plenty that needs to be done" that marks, for me, when Turing's statements may have started to become really prone to ambivalence with respect to his Promethean ambition. And it has only got worse in *c.* 1951, when — perhaps now even more aware of the risks and decided to be a prophet of the machines — he subentitled his BBC radio lecture "a heretical theory" and said:

> Let us now assume, for the sake of argument, that these machines are a genuine possibility, and look at the consequences of constructing them. To do so would of course meet with great opposition, unless we have advanced greatly in religious toleration from the days of Galileo. There would be great opposition from the intellectuals who were afraid of being put out of a job. It is probable though that the intellectuals would be mistaken about this. There would be plenty to do, trying to understand what the machines were trying to say, i.e. in trying to keep one's intelligence up to the standard set by the machines, for it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control, in the way that is mentioned in Samuel Butler's 'Erewhon'. (TURING, 2004 [*c.* 1951], p. 475)

Turing thus reasoned. If intelligent machines are a genuine possibility — and he had already positioned that he thought so — then it seemed probable that by force of the machine thinking method machines would soon outperform our intelligence powers, and eventually take control. Turing had already included in his (1950) bibliography Samuel Butler's "Book of the machines" which composes three chapters of *Erewhon*, but no direct reference to it can be found in the text itself. My guess is that it refers to the passage "[t]hese are possibilities of the near future, rather than Utopian dreams" (p. 449) appearing in the end of his discussion of the fifth objection, "arguments from various disabilities." So in the above *c.* 1951 passage he cited Butler's famous novel for the second time, then much more explicitly and in clear association with his own views.

Now, one may wonder how did Turing scholars read passages such as the above. We have seen Turing biographer Andrew Hodges to side with Jefferson's view of Turing's Promethean ambition (§1.1). Jack Copeland in his turn declared:

> Turing ends 'Intelligent Machinery, A Heretical Theory' with a vision of the future, now hackneyed, in which intelligent computers 'outstrip our feeble powers' and 'take control'. There is more of the same in [Turing's 2004 [1951]]. No doubt this is comic-strip stuff. (COPELAND, 2004, p. 470)

More recently, Diane Proudfoot also seems to shrink the seriousness behind Turing's irony:

> Turing (following Butler) poked fun at the fear of out-of-control AI. When he predicted in the London Times that machines could "enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms," the media protested at the "horrific" implication of these ideas — namely, "machines rising against their creator." But Turing said drily, "A similar danger and humiliation threatens us from the possibility that we might be superseded by the pig or the rat." He joked — but with a good pinch of his usual common sense — that we might be able to "keep the machines in a subservient position, for instance by turning off the power at strategic moments." (PROUDFOOT, 2015)

But Proudfoot would add to that seemingly to concede for some seriousness in Turing's prophecy:

> Turing's response to AI panic was gentle mockery. All the same, there was a serious edge to his humor. If runaway AI comes, he said, "we should, as a species, feel greatly humbled." He seemed almost to welcome the possibility of this humiliating lesson for the human race. (PROUDFOOT, 2015)

I am not sure whether Proudfoot granted to Turing the view that the possibility of having machines superseding us was real or just a joke about our species chauvinism. It is common to Hodges, Copeland and Proudfoot's references altogether, in any case, that they only refer to what I have called the problem of identifying Turing's specific Promethean ambition in side notes (Hodges and Copeland) and web blog posts (Proudfoot). The same holds for their comments on Butler's *Erewhon*, if any. I think, nonetheless, as suggested early on in this chapter, that the problem does have importance and deserves study.

Accordingly, I shall now take a moment to introduce Butler's novel and a key source about how it reached and may have influenced Turing.

Turing, reader of Samuel Butler's *Erewhon* (1872)

Butler's *Erewhon* (1872) became popular in Victorian England early on since it was published. According to Mrs. Sara Turing, it would have reached Turing early in his youth in the turn from the 1920's to the 1930's. As she recollected, its influence on Turing may have been quite deep:

> In [Alan's] late teens he read a certain amount of fiction [...]. He had a particular fondness for The Pickwick Papers, George Borrow's books and Samuel Butler's Erewhon. This last possibly set him to think about the construction of an actual intelligent machine. (TURING, 2012 [1959], p. 108).

Butler's novel can be considered a social critique in the form of a satire, hard to class either as utopia or dystopia, as things are presented always from two perspectives. Butler lived in the (Victorian) machine age, while the philosophical and cultural shockwaves of Charles Darwin's 1859 *On the origin of species* were still being felt. Butler thus applied in a non-obvious way natural selection to machinery itself. Butler had first published, in 1863, in Christchurch newspaper *The Press*, "Darwin among the machines."[2] That would be revised and enlarged to become in 1872 "The book of the machines," which is itself a fictional book within *Erewhon* (the fiction), and its description by the narrator takes three chapters of *Erewhon*, as just mentioned. "Erewhon" is "nowhere" written backwards (or almost), and referred to a yet undiscovered country in New Zealand. There, so runs the novel, a revolution was started and led to civil war, opposing "machinists" and "anti-machinists." It was won by the latter. Then all of the more complicated machines formerly in common use were destroyed, and all treatises on mechanics burned (1872, p. 188). Only a few hundred years later, thus the story goes, when no Erewhonian seemed to consider anymore the idea of reintroducing forbidden inventions, the subject came to be regarded as a curious antiquarian study. So this is how the book of the machines gets written within the novel, as a recollection of the arguments of anti-maquinist philosophers. Here is an example addressing the rate of progress of machines or of the so-called "mechanical kingdom:"

> Reflect upon the extraordinary advance which the machines have made during the last few hundred years, and observe how slowly the animal and vegetable kingdoms are advancing in comparison. The more highly organised machines are creatures not so much of yesterday as of the last five minutes, so to speak in comparison with past time. (BUTLER, 1872, p. 191).

Butler's anti-maquinist arguments also comprised aspects such as the emergence of consciousness and a purely mechanical reproductive system. That was all reasonable for Butler to write about in light of the industrial revolution and Darwin's then new theory of evolution by natural selection. And it must have seemed persuasive to Turing indeed, who would in (1936) conceive of a universal machine and ever since be a key figure to get the computer revolution underway.

We shall now be ready to examine in more depth Turing's 1950 and *c.* 1951 citations of Butler, and they did not come out of nothing.

[2] Reproduced in the New Zealand Electronic Text Collection at <http://nzetc.victoria.ac.nz/tm/scholarly/tei-ButFir-t1-g1-t1-g1-t4-body.html>. Access on 16 Jul 2020.

On Turing's citations of Butler's *Erewhon*

Butler's *Erewhon* had actually been cited first by Norbert Wiener in his *Cybernetics* (1965 [1948], p. 27), a book that received commentary both from Jefferson in his (1949a) Lister oration and from the BMJ's 1949 editorial that I quoted before (§1.3). In fact, after having admonished Turing directly this BMJ's editorial mentioned *Erewhon* and warned: "if we fail to recognize [that] the mind must surely be greater than its own ideas about itself, [...] we may suffer the fate of the [Erewhonians] and be enslaved by machines [...]" (BMJ, 1949, p. 1130). This was a prototype of the kind of opposition that Turing faced. One may observe in the conditional statement that the antecedent — "the mind must surely be greater than its own ideas about itself" — engages in a petition of principle. It was as if the theoretical and empirical possibility of (super)intelligent machines, on the one side, and the real status of the human mind, on the other side, were both just a matter of morals, crime and punishment. All this happened in June 1949.

Considering Turing's specific *c.* 1951's citation of Butler as quoted above, chances are that Turing did not forget the BMJ's editorial. And it seems that he wanted to shock. Turing, who had already some appreciation for Butler's fiction himself, seems to have found an opportunity to respond ironically to it. He used irony to imitate his contenders' language and render a scenario that was the exact opposite of what they considered desirable. And yet this was not at all an empty joke of his, for as we shall see next, the evidence is suggestive that he did think that the possibility that machines may outstrip our intellectual powers was both true and reasonable. I interpret that Turing's *c.* 1951 citation of Butler carried a message. It was not parody but *satire*.

It is actually recurrent in Turing sources the advent of superintelligent machines, sometimes with irony, sometimes without. I now proceed to present representative instances. A key such statement, and to my knowledge the first one chronologically, is Turing's 1948 formulation of objection "(a)" which I quoted before (§1.2). Turing's tone there was steady, and that source is in fact an official NPL report. Two other most notable statements are the following. First, from a vivid 1949 reminiscence of Lyn Irvine's, we learn about this Newman-Turing dialogue:

> I remember sitting in our garden at Bowdon about 1949 while Alan [Turing] and my husband [Max Newman] discussed the machine ('Madam') and its future activities. I couldn't take part in the discussion and it was one of many that had passed over my head, but suddenly my ear picked up a remark which sent a shiver down my back. Alan said, reflectively, 'I suppose when it gets to that stage we shan't know how it does it.' (TURING, 2012 [1959], p. 95)

I take to be fairly clear that Turing's observation as reported above was not ironic. First, one may consider that this was a private conversation between two long-standing collaborators and friends, whose positions about machine intelligence did not differ enough to justify use of irony there. Second, because Irvine herself gave away the tone of Turing's statement, as she said "suddenly my ear picked up a remark which sent a shiver down my back." Now, while Turing did

not mention machines outstripping our powers and taking control specifically, I think that the context, most notably the implied tone that sounded dreadful to Irvine, adds up to make it.

For a second related statement of Turing's, let us consider his May 1951 BBC radio lecture (§A.4.4). In the two last paragraphs of the lecture's transcript, Turing revisited once more what boils down to "the heads in the sand" objection, now perhaps in a more diligent tone, say, more like in (2004 [1948]), and less like in (2004 [*c.* 1951]). He again recollected the theoretical possibility of other species, say, "the pig or the rat," to supersede us humans in intelligence power, and compared it with the same possibility now applied to the machines. He then added:

> But this new danger is much closer. If it comes at all it will almost certainly be within the next millennium. It is remote but not astronomically remote, and is certainly something which can give us anxiety. (TURING, 2004 [1951], p. 486)

Turing thus plainly posed: the advent of superintelligent machines is not a certain event, but its possibility is genuine. It will take a long time, but it may come in a foreseeable future. Now, if he thought so and still wanted to build intelligent machines, does that mean that he actually *wanted* machines to take control? I think not. As anticipated, I think that Turing was, among other things, raising a precautionary voice.

In the presence of the crass chauvinism of some of his contenders, however, he could not but establish epistemological relations between the possibility of *intelligent* machines and the refutation of what he saw as species and gender biases of his contenders. This is, I think, the reason why Turing's prophesized future pervaded by intelligent machines may be seen as a dystopia just as much as a utopia. It depends on the social and cultural background of who is seeing. *Superintelligent* machines were for him a true but distant possibility that pushed his argument to the limit from both epistemological and social stances.

I shall now proceed to consolidate this analysis of the problem of identifying Turing's specific Promethean ambition.

Turing's Promethean ambition

I find in the system of Turing's propositions that when he discusses the possibility of machines to supersede us in intelligence power, what he meant is that it is a serious problem that needs to be studied in detail, both analytically and empirically. In fact, there is a key aspect towards resolving the question which is his reference to the timescale aspect, in particular, his dismissal of the urgency of time. Turing considered the advent of superintelligent machines to be not astronomically remote yet remote enough to allow for deeper studies with the prospect of giving us clear benefits in the meantime. In the same 1951 lecture, Turing followed on and admitted:

> The whole thinking process is still rather mysterious to us, but I believe that the attempt to make a thinking machine will help us greatly in finding out how we think ourselves. (TURING, 2004 [1951], p. 486)

This passage is not filled with irony either. It is perhaps the most direct statement that Turing made about his deeper positive motives. I take it to be confessional about his hopes: they were pointed towards improving our scientific understanding of the human mindbrain. There is also secondary-source evidence of that from a testimony by Mrs. Sara Turing. She reports that Turing has told her about his aims as early as 1944:

> Sometime round about 1944 he [Turing] had talked to me about his plans for the construction of a universal computer and of the service such a machine might render to psychology in the study of the human brain. This he regarded as likely to be one of the more valuable contributions a universal computing machine could make to knowledge. (TURING, 2012 [1959], p. 92)

Given this secondary source and what we have seen so far, I take that Turing's positive aims were laid out in view of human enlightenment broadly construed. We may now ask, are these aims of Turing's actually new in history? Do they have a strain? I think yes, they do.

Turing can be associated with the modern tradition of mechanistic natural philosophy. Besides his connections with Descartes as we shall examine in detail later (§3), he can be approximated, for example, with figures such as the French inventor and engineer Jacques de Vaucanson (1709-1782), famous for, among other pieces, his artificial duck. The designers of early modern automata from the seventeenth and the eighteenth centuries have often been criticized on the grounds of being driven by futile, entertaining-only motives. Is this fair? "No," David Fryer and John Marshall suggested in (1979). They presented an interesting study of the motives of Vaucanson. The "claim that the primary objective of Vaucanson's work was 'to astonish and amuse the public'," Fryer and John Marshall posited, "is hardly fair" (p. 267-8). "Vaucanson," they concluded, "was an entertainer, but he was also deeply committed to the development of an explanatory psychology." Fryer and Marshall's findings about Vaucanson seems to offer a promising path to locate Turing's motives. But Turing was definitely no entertainer. Their inquiry, nonetheless, is within a class of studies presented earlier by Silvio Bedini (1964) and Derek de Solla Price (1964), which showed that early modern automata were neither "trivial toys" nor "immediately useful inventions." Rather, they were simulacra or models "whose very existence offered tangible proof, more impressive than any theory, that the natural universe of physics and biology was susceptible to mechanistic explication" (1964, p. 9). This seems quite accurate as a description of the Turing's most primary aims. And yet, unlike the court natural philosophers of the seventeenth and eighteenth centuries, Turing eventually turned to be, as we have seen, rather a *social critic*. We shall now have a firmer basis to identify Turing's Promethean ambition. The most explicit aspect of Jefferson's view — which defined Turing as "a sort of scientific [Percy] Shelley" (2012 [1959], p. 58) — seems just about right. In light of all that we have seen, I interpret, Percy Shelley's enlightened Prometheus suits Turing.

I would like to pose interpretive boundaries of a negative kind as well. I shall discourage the association of Turing's ambitions with the image of Mary Shelley's Dr. Frankenstein. For

instance, as we have seen, Hodges offered a view of Turing as "a Frankenstein — the proud irresponsibility of pure science, concentrated in a single person" (2012 [1983]), p. 521). But Frankenstein is widely seen as the figure of a scientist who is blind in the pursuit of his science and unwary of its consequences. So a question that I pose to this view is the following. If Turing impersonated Dr. Frankenstein, why would he have exposed himself and spent time raising a voice in the public domain about the possibility of having machines outstripping our intellectual powers and taking control over us several decades, perhaps more than a century before he thought this possibility was feasible? It does not seem to fit. In fact, juxtaposed to the passage quoted above when he positively stated his aim of finding out how we think, Turing gave this negative statement:

> But I certainly hope and believe that no great efforts will be put into making machines with the most distinctively human, but non-intellectual characteristics such as the shape of the human body; it appears to me to be quite futile to make such attempts and their results would have something like the unpleasant quality of artificial flowers. Attempts to produce a thinking machine seem to me to be in a different category. (TURING, 2004 [1951], p. 486)

The passage offers evidence that Turing did *not* find interesting to put effort into making fantastic things such as, say, an artificial creature resembling the human body, and even thought this to be futile and bound to give unpleasant results. Viewing Turing as a sort of Dr. Frankenstein — say, in connection with Hodges' suggestion and with Jefferson's dubious association of him with (the) Shelley(s) and with Mays' allusion to a "mechanical necromancer" — is unsupported. Furthermore, with reference to the "unpleasant quality of artificial flowers," Turing expressed his dislike to the dismissal of the ontological distinction between the natural and the artificial. So the evidence is unfavorable to associating Turing with, say, modern transhumanist movements.

Turing, as we have seen, was first and foremost a scientist. He believed in human rationality and was not afraid about getting to establish our own limits as human beings. He believed that he could be able to pursue further mathematical studies to anticipate the logical possibilities and limits of machines, and that this was actually needed given the possibility of superintelligent machines. He wanted to study the status of machine intelligence as the other face of our own status as human beings. This was not a fantastic project, but one laid out towards human enlightenment. The evidence is in favor of viewing Turing as a humanist, and not as a transhumanist. Challenged by a conservative reaction, Turing did not pass on the opportunity to make a subtle social critique about our species and gender chauvinisms, and even racial and national partisanships altogether. It is precisely because he did *not* think that we humans are necessarily superior beings, that associating Turing with the arrogance that concerned Mary Shelley as portrayed in the character of Dr. Frankenstein, is just far-fetched.

1.6 Turing's plea

"Fair play for the machines" when testing their intelligence, it was all that Turing asked in the formulation of his "plea." In his February (2004 [1947], p. 394) lecture, he said:

> Against [the argument based on a result from mathematical logic that a machine, but not a human mathematician, will in some cases fail to give an answer] I would say that fair play must be given to the machine. Instead of it sometimes giving no answer we could arrange that it gives occasional wrong answers. [...] the human mathematician would likewise make blunders when trying out new techniques. It is easy for us to [...] give him another chance, but the machine would probably be allowed no mercy. To continue *my plea for 'fair play for the machines'* when testing their I.Q. [,...] the machine must be allowed to have contact with human beings in order that it may adapt itself to their standards. (TURING, 2004 [1947], p. 394, emphasis added)

This must have been a high moment in Turing's lecture to the London Mathematical Society. That lecture came in the wake of the 1946 discussions in the British press in relation to whether or not it would make sense to refer to electronic computing machines as "electronic brains" (§A.3.3). So, as of early 1947, there was already some opposition to the views that Turing wanted to uphold. And, in effect, in the 1947 lecture we see Turing starting to articulate rebuttals to two important objections that later in (1950) he would name "Lady Lovelace's objection" and "the mathematical objection." In order for us to better understand Turing's plea, let us briefly review Turing's approach relative to the objections to machine thinking.

Whose side is the burden of proof on?

Turing took over one third of his 1950 paper in rebuttal to those nine objections. As mentioned, he thought that he did not have very convincing arguments of a positive nature to support his views. He said: "[i]f I had, I should not have taken such pains to point out the fallacies in contrary views" (p. 454). Implicit in Turing's rationale, there were two aspects I want to highlight. First, the original question, whether machines can think, could not altogether be abandoned, "for opinions will differ as to the appropriateness of the substitution and we must at least listen to what has to be said in this connection" (p. 442). Second, as we have seen before (§1.4), for Turing "doing the [imitation game] experiment" was the "only satisfactory support" and yet it was actually unfeasible back then due to the lack of the required storage capacity. He thus had to engage in the philosophical discussions. And over all of Turing's argumentation there is a distinguishing mark, I propose, which is his plea for fair play for the machines. Turing's obituary in *The Times*, for example, captured that by saying: "his view expressed with great force and wit, was that it was for those who saw an unbridgeable gap between the two [the human mindbrain and the digital computer] to say just where the difference lay" (1954).

Indeed, against most of the lines of attack to his views on machine thinking, Turing pushed back by shifting the burden of proof towards human thinking and the human mind. I

shall refer to that as Turing's plea for fair play for the machines. In (2004 [1948], p. 410-2), Turing first dealt more explicitly with what he saw to be the main objections. He then formulated and rebutted five objections, labeled in letters *a* to *e*. Later, in (1950, p. 443-54), Turing revised and enlarged them to nine named objections. In five out of the nine objections, I hold, Turing's argumentation can be best understood as a shift in the burden of proof. I present in Fig. 1 a scheme relating the 1948 to the 1950 objections and marking those that, in my interpretation, Turing replied by shifting the burden of proof. I indicate below passages from Turing sources in support of it. I will first quote Turing's formulation of the objection, and then quote his reply.

- *The theological objection.* "Thinking is a function of man's immortal soul. God has given an immortal soul to every man and woman, but not to any other animal or to machines. Hence no animal or machine can think" (1950, p. 443). Turing's reply: "In attempting to construct such machines we should not be irreverently usurping His power of creating souls, any more than we are in the procreation of children: rather we are, in either case, instruments of His will providing mansions for the souls that He creates" (*Ibid.*, p. 443). So, Turing argued, if building an intelligent machine is an irreverent usurpation of God's power, then one should prove why human procreation is not.

- *The mathematical objection.* "If [a machine] is rigged up to give answers to questions as in the imitation game, there will be some questions to which it will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply" (1950, p. 444). Reply: "although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect" (*Ibid.*, p. 445). So, Turing argued, if Gödel's and related results impose limitations to the power of any particular machine, then one should prove why the same limitations do not hold for any particular human being.

- *The argument from consciousness.* "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain — that is, not only write it but know that it had written it" (1950, p. 445). Reply: "According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking. [...] Likewise according to this view the only way to know that a *man* thinks is to be that particular man. It is in fact the solipsist point of view" (*Ibid.*, p. 446-7, no emphasis added). So, Turing argued, if accepting that a machine thinks requires knowing that it has introspective feelings, then one should prove how exactly they know that any particular fellow human being has such feelings.

- *Lady Lovelace's objection.* "In so far as a machine can show intelligence this is to be regarded as nothing but a reflection of the intelligence of its creator" (2004 [1948], p. 411; later rephrased in 1950, p. 450). Reply: "The view (e) that intelligence in machinery is

merely a reflection of that of its creator is rather similar to the view that the credit for the discoveries of a pupil should be given to his teacher" (2004 [1948], p. 411; later rephrased in 1950, p. 450, and p. 458-9). So, Turing argued, if a machine can never do anything really new with respect to what it has been instructed by its designer, then one should prove how what a (human) student or pupil does is "not simply the growth of the seed planted in him by teaching, or the effect of following well-known general principles" (1950, p. 455).

- *The argument from informality of behavior.* "To attempt to provide rules of conduct to cover every eventuality, even those arising from traffic lights, appears to be impossible. With all this I agree. From this it is argued that we cannot be machines. I shall try to reproduce the argument [...]. 'If each man had a definite set of rules of conduct by which be regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines.' The undistributed middle is glaring" (1950, p. 452). Reply: "There may however be a certain confusion between 'rules of conduct' and 'laws of behaviour' to cloud the issue. By 'rules of conduct' I mean precepts such as 'Stop if you see red lights', on which one can act, and of which one can be conscious. By 'laws of behaviour' I mean laws of nature as applied to a man's body such as 'if you pinch him he will squeak.' If we substitute 'laws of behaviour which regulate his life' for 'laws of conduct by which he regulates his life' in the argument quoted the undistributed middle is no longer insuperable. [...W]e cannot so easily convince ourselves of the absence of complete laws of behaviour [...]. The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, 'We have searched enough. There are no such laws'." (*Ibid.*, p. 452). So, Turing argued, in order to prove that men cannot be machines, one would have to rely on scientific observation to search for laws of behavior — not for rules of conduct which one can be conscious of — and, in any case, trying to establish the negative result would engage in evasive scientific induction.

Now, if Turing replied to five out of nine of his 1950 objections by shifting the burden of proof as part of his plea for fair play, does that mean that he was evading the responsibilities that should come together with his claims? I think not. Turing had already given in (2004 [1948]) the core reason why shifting the burden of proof was, in this case, indispensable in a positive sense too. For him, "the idea of 'intelligence' is itself emotional rather than mathematical" (p. 411). (We shall see that in detail later in §2.3.) Therefore, it is only by pushing back the burden of proof that he would be able to clear the ground for the reception of his views. In fact, arguing for the possibility of machine thinking seemed to require reviewing what was meant by (human) thinking in the first place, and this could lead to reviewing the ontological status of human. Specifically, it could challenge a certain notion of free will. (Cf. above Turing's reply to the argument from informality of behavior.) I want to promptly suggest that this is one of the core reasons why Turing's views have been received in general as inconvenient, as I discuss next.

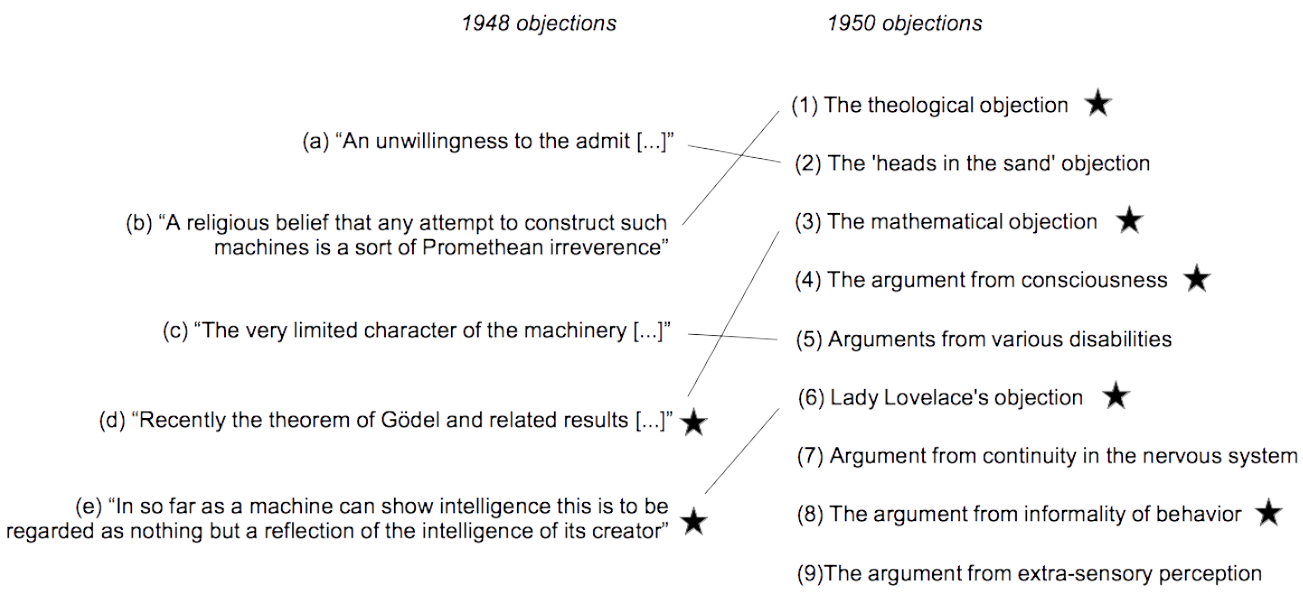

Figure 1 – The 1948 and 1950 objections formulated and rebutted by Turing. Starred objections are those which in my interpretation Turing rebutted by shifting the burden of proof towards human thinking. I shall refer to this scheme from further chapters as well.

Specific *ad hominem* arguments in Cold War Britain

In 1949, Jefferson said: "the concept of thinking like machines lends itself to certain political dogmas inimical to man's happiness [and] erodes religious beliefs that have been mainstays of social conduct." This was a moment of his (1949a, p. 1107) Lister oration, which was published in the 25 June issue of the *British Medical Journal*. The BMJ editorial, which cited Turing directly (§1.3), has seemingly been less subtle than Jefferson:

> There is an undeniable danger in the facile acceptance of materialism, for the materialist finds values and ethics an insoluble problem. If the mind is ultimately mechanical, what is the source of standards of right and wrong? These can be only the result of individual and social conditioning and can have no more than personal validity. Wiener, to whose book *Cybernetics* Jefferson alludes, explains that this is done by setting three machines the same problem and, if they do not all agree on the answer, accepting the verdict of the majority. [...] Inevitably in the world of machines ethics is decided by the majority. This generation, at least, knows where that doctrine leads — to tyranny, the concentration camp, the gas chamber, and the cremation oven. (BMJ, 1949, p. 1130)

The association of the idea of intelligent machines with materialism and nazism may strike the reader these days. It was back then suggested by several antagonists including, for example, computer pioneer Douglas Hartree (Cf. HODGES, 2012 [1983], p. 348). In fact we may approximate this phenomenon to a specific form of *ad hominem* argument, which has been coined *reductio ad Hitlerum* by Leo Strauss in 1953 (and I cite this, *n.b.*, because the publication date adds specific historical evidence in connection with the attacks on Wiener and Turing):

> Unfortunately, it does not go without saying that in our examination we must avoid the fallacy that in the last decades has frequently been used as a substitute for the *reductio ad absurdum*: the *reductio ad Hitlerum*. A view is not refuted by the fact that it happens to have been shared by Hitler.
> (STRAUSS, 1999 [1953], p. 42, no emphasis added)

The logical and political fallacy runs this way:[3] Adolf Hitler or the Nazis accepted idea *i*; person *p* or group *g* accepts *i*; therefore, *p* or *g* is as bad as Hitler or the Nazis. It also became known as playing the Hitler card or the Nazi card. The specific views of Wiener and Turing that were claimed by contenders to have been shared by Hitler or the Nazis may vary in each such attack.

In fact, the Nazi card played by a BMJ editorial against Wiener and Turing's views can actually be seen from a broader perspective. There was a social and cultural campaign on "freedom" against "totalitarianism" running back then (ROMERO, 2017, p. 291). Take for a memorable related milestone, say, Wiston Churchill's *The Sinews of Peace* ("Iron Curtain") speech at Fulton, Missouri, on 5 March (1946). This is all worth mentioning as it shall inform the historical context of the reception of Turing's test in the early 1950's (§3.2). It turns out that Turing was presenting his bold views on intelligent machines against the background of Postwar, Cold War Britain. This period has been called "Second Red Scare" (1947-1957), and sometimes "McCarthyism," in relation to Cold War hysteria both in the US and in Britain. It was fashionable to raise a voice on human freedom back then. As we have seen (§1.3), Jefferson had referred to Turing as "non-conformist" (TURING, 2012 [1959]), p. 58). In fact, that was a common designation in the early 1950's to allude for suspicion of lack of patriotism and possible involvement with Communism, both in the US and in Britain. For example, British citizen Charles Chaplin was forbidden in September 1952 to enter in the US based on charges of political and moral offenses (HICKEY, 1969). Chaplin would write later in *My Autobiography* (1966 [1964]): "[f]riends have asked how I came to engender this American antagonism." He answered: "[m]y prodigious sin was, and still is, being a non-conformist."

It is unlikely that Turing's contenders, including Jefferson and Mays, knew back then that his views on intelligent machines, whether (as implied by Wolfe Mays) a sort of witchcraft or not, had roots in his contributions to actually save their country from the threat of Nazism. For a fact, take that in June 1946 Turing was appointed an Officer of the Order of the British Empire (OBE) by King George VI for his wartime services, but his work remained secret for many years. The biographies of Hodges (2012 [1983]) and Copeland (2012) gathered information enough to support classing Turing as patriotic and friend of human freedom. After having delivered extraordinary service to his country for over thirteen years (ever since September 1938), though, he would be officially charged in early 1952 of offending the customs by his own National State.

[3] I benefited from Gary N. Curtis' page: <http://www.fallacyfiles.org/adnazium.html>. Access on 26 July 2020.

1.7 Turing's fate

Turing was gay. Near Christmas 1951, he would have met a young man, Arnold Murray, and started with him some relationship that went through January 1952 (HODGES, 2012 [1983]; COPELAND, 2017, p. 449; p. 36). On 23 January, Hodges reports (p. 454), Turing's house was burgled and he would have heard from Murray an apparently strong clue about the burglar, an acquainted with Murray. For Hodges, it was perhaps by being afraid of blackmail in relation to his identity as a homosexual that Turing proceeded to report the crime to the police (p. 454-5). In any case, Copeland related, during questioning Turing gave away three times his sexual relationship with Murray (p. 36). A case was formed under "Gross Indecency contrary to Section 11 of the Criminal Law Amendment Act 1885." Hodges reported the detectives to have said: "[h]e was a real convert [...] he really believed he was doing the right thing" (p. 457).

Let us now pause and look for a moment more broadly at Turing's public image. As if having talked three times at the BBC from 1951 to 1952 about the possibility of machines to supersede us in intelligence and taking control were not enough, now there was a second shade of non-conformism to be attached to him. The case "The Queen versus Alan Mathison Turing" attracted attention of the local newspapers, whose archives have been obtained by Copeland (p. 36, note 3). Turing was brought to trial on 31 March 1952. Hodges related (p. 472) that Max Newman and Hugh Alexander (Turing's peer in his Foreign Office service during World War II), having testified as witnesses in the trial, were impressed by Turing's "strong line" (Newman) and "moral courage" (Alexander). He was convicted and submitted to one year of chemical castration by injection of female hormones. (Today, conversion therapy, also known as the pseudo "gay cure," can be considered as a form of torture and is considered potentially harmful by major therapy professional bodies in Britain and worldwide.

Now, how does all that relate with Turing's views on machine intelligence and the role he had assumed as prophet of the machines? It turns out that several weeks before the trial, on 10 January 1952, Turing participated in a round table at BBC to debate point and counterpoint with Jefferson, having Max Newman and Cambridge-philosopher Richard Braithwaite as sort of referees. The discussion was aired first on 14 January, and again on 23 January. We shall come back to it in some detail later (§A.4.6). My goal now is to reflect on how Turing assimilated it later in a letter to his friend Norman Routledge in *c.* early 1952. Here is a large excerpt of it:

> I am not at present in a state in which I am able to concentrate well, for reasons explained in the next paragraph.
>
> I've now got myself into the kind of trouble that I have always considered to be quite a possibility for me, though I have usually rated it at about 10:1 against. I shall shortly be pleading guilty to a charge of sexual offences with a young man. The story of how it all came to be found out is a long and fascinating one, which I shall have to make into a short story one day, but haven't the time to tell you now. No doubt I shall emerge from it all a different man, but quite who I've not found out.

> Glad you enjoyed broadcast. Jefferson certainly was rather disappointing though. I'm afraid that the following syllogism may be used by some in the future.
>
> Turing believes machines think
> Turing lies with men
> Therefore machines do not think
>
> Yours in distress, Alan. (TURING, 2012 [*c.* early 1952])[4]

Biographer David Leavitt observed that "to lie with" is a biblical locution (2006, p. 5), whose use by Turing is unlikely to have been accidental. In the letter, as one may observe, Turing expressed disappointment and took the sexual-offences charges and/or their consequences to be a milestone in his life. Yet mostly interesting to my focus here is the connection that Turing established between those events and the polemic with Jefferson. If he could hold tight in face of *ad hominem* attacks since June 1949, it seems that they would now come to be, for him, overwhelming and unmanageable. The social embarrassment and exposure would now be too great and sit in obstruction to the philosophical dispute. Thus the public accusation about his homosexuality and punishment by chemical castration translated into social castration with respect to his philosophical positions. Aside from a *c.* 1952 text where Turing restated timidly a couple points on machine intelligence (§A.4.8), the 1952 BBC broadcast was the latest communication of Turing specifically on that topic. Back in the highs of Turing's defense of mechanical intelligence in (2004 [*c.* 1951]), Turing said that to assume that intelligent machines are a genuine possibility "would of course meet with great opposition, unless we have advanced greatly in religious toleration from the days of Galileo" (p. 475). Soon after then, in effect as we have seen in his *c.* early 1952 letter to Routledge, Turing would have found an ultimate response: such advances had not yet come by.

In (2013), Daniel Dennett proposed an analogy between Turing and the great natural historian, Charles Darwin. He referred to Turing and Darwin's "strange inversion of reasoning," and said "[w]hat Darwin and Turing had both discovered, in their different ways, was the existence of *competence without comprehension*" (p. 571, no emphasis added).

I rather would like to suggest a connection of Turing with Galileo Galilei. This was a figure whom Turing himself liked to refer to, and I shall argue, not without reason. For a specific association, I shall identify in Chapter 3 a remarkable resemblance of Turing to Galileo in their uses of thought experiment. I present now a general association. Galileo faced opposition in Renaissance Italy because of his views on whether the Earth moves, which had deep consequences to our world view. Galileo was sent to trial, found guilty and convicted to "formal imprisonment," which on the condition that he abjured his heresies was changed to house arrest (FINOCCHIARO, 1989). Galileo was imposed "[a]s a salutary penance [...] to recite the seven penitential Psalms once a week for the next three years" (p. 291). Turing faced opposition in post-war Britain because of his views on intelligent machines, which promised to have deep

[4] This letter has been read by actor Benedict Cumberbatch. Available at: <http://www.youtube.com/watch?v=57qs8Y_aqcc>. Access on 20 July 2020.

consequences to our view of ourselves and, in particular, to a certain notion of free will. (I shall discuss the topic of free will in connection with Turing's thought later in §2.3). Although not because of his scientific and philosophical heresies, he was also sent to trial and had to plead guilty. He was convicted to a sentence comparable to house arrest, and had to inject in himself female hormones. Once submitted to that, he felt somehow prevented to keep up holding his heretical views on intelligent machines. In between his 1616 and 1633 trials, in his 1623 *Assayer*, Galileo wrote:

> I believe that they [the good philosophers] fly, and that they fly alone, like eagles, and not in flocks like starlings. It is true that because eagles are rare birds they are little seen and less heard, while birds that fly like starlings fill the sky with shrieks and cries [...]. (GALILEI, 1957 [1623], p. 237)

Given all that we have seen so far, I think that Galileo's eagles-starlings metaphor fits very well to Turing's biography. I shall just add that Turing, given the image implied by his iconic "heads in the sand" objection, might have liked better to allude to (a flock of) ostriches instead.

Turing's predecessors in the modern history of computing, say, Gottfried W. Leibniz, Blaise Pascal and Charles Babbage, were all (in their own specific ways, of course) devoted to God. Babbage (1791-1871), in particular, found inspiration in his specific natural-theology figure of God (a divine programmer of the universe) for the design of his Analytical Engine (2009 [1837]). Turing, instead, found inspiration in our own minds by observing the intellectual behavior of calculation clerks, the human computers. This marks a relevant distinction in how we shall situate Turing in the history of philosophy. Also, we may once more recall the way Newman has put it: "[t]he central problem with which [Turing] started, and to which he constantly returned, is the extent and the limitations of mechanistic explanations of nature" (1955, p. 256). In fact, human thinking and/or the human mind, specifically, were Turing's intellectual target.

Since the serious difficulties of early 1952, it seems that from mid 1952 on Turing was getting back on track to his good spirits. In May 1952, he attended a meeting of the Ratio Club (the British circle on cybernetics) held in Cambridge (HUSBANDS; HOLLAND, 2008, p. 124), and was photographed with others in apparently good mood (cf. *Ibid.*, Fig. 6.2). In October 1952, Hodges reports, Turing would have begun to see a Jungian psychoanalyst, Franz Greenbaum (2012 [1983], p. 480). In the same month — back in connection with the intellectual community with which Turing engaged in discussions on machine thinking from 1949 to early 1952 —, Turing would have attended a course of lectures given by Jean Piaget at the University of Manchester Philosophy Department (*Ibid.*). His interest in Piaget must have been related to the problem of educating a child machine as he had thematized most notably in 1950 and 1951 (§A.4). One may then wonder that he could have been able to eventually get back to the public stage in relation to his research on the human mind.

Turing was found dead of cyanide poisoning in 8 June 1954 just before completing 42 years old. The three possibilities to explain the poisoning are suicide, accidental inhalation, and

assassination (COPELAND, 2012, p. 226). The official version was suicide. Turing would have bitten an apple laced with cyanide. But the case for this seems to be weak. Although a bitten apple was found near Turing's bed, it was never tested for cyanide; not to mention other facts that do not match very well (*Ibid.*, p. 223-4). And yet that version, which was never accepted by Mrs. Sara Turing (1881-1976) herself, accompanies even her own biography, whose centenary edition (2012 [1959], posthumous to her) is stamped with a bitten apple on the front cover.

The announcement of Turing's royal pardon came in 2013 (BBC), after an increasing social pressure led by British scientists and intellectuals — a first petition to the British government was rejected in the occasion of Turing's centenary year in 2012 (BBC). In (2009), yet before the Queen's forgiveness, Prime Minister Gordon Brown had issued a public apology. Among the PM's words, *The Guardian*'s Caroline Davies featured in her headline his saying: "we were inhumane." I take this note, issued over 50 years later, as sufficiently symbolic in contrast to how Turing was seen by Jefferson and others as of the early and mid 1950's.

Turing believed that machines can think, and that machines will think. I have tried to show that he stood up for announcing to the world that there was a need to take the capabilities of (future) machines seriously. Erich Fromm, in his 1967 contribution to *Bertrand Russell, philosopher of the century*, wrote:

> Those who announce ideas — and not necessarily new ones — and at the same time live them we may call *prophets*. [...] It is not that a prophet wishes to be a prophet; in fact, only the false ones have the ambition to become prophets. (FROMM, 2010 [1967], p. 14-5, no emphasis added)

I hope to have succeeded in my assembly of Turing as a true prophet of the machines.

1.8 Analytical summary

In this chapter I have presented an intellectual-biography sketch of Alan Turing (1912-1954) with focus on the role he assumed in the public controversy that took place in England (1949-1952) on whether machines can think. In particular, I have studied Jefferson's suggestion that Turing was "a sort of scientific Shelley." I now offer a summary collecting my findings and the key points hereby developed.

Irreverence (§1.2). Turing engaged in what contenders considered to be, in his eyes, a sort of Promethean irreverence. While having cultivated friendship with several fellow mathematicians and scientists who learned to respect and much admire him, he was seen by those contenders as if he had lived in a "slightly inhumane world," or as a "mechanical necromancer." I have suggested that his contenders were slippery between two conflicting understandings of his views, which are labeled interpretation (i) and (ii) in the end of §1.2. They suggested, sometimes that (i) his views were feasible but bound to have dangerous social and ethical implications, and

sometimes that (ii) they were unfeasible, fantastic, or born out of some sort of wishful thinking and bound to do damage by misleading the public opinion.

Irony (§1.3). I have presented spontaneous testimonies from Turing's friends concerning his sense of humor. They relate that Turing had modesty and intellectual integrity, that he was able to keep remarkable balance between humor and seriousness. In connection to the philosophical dispute on intelligent machines, he received *ad hominem* attacks and held tight. In meeting with great intellectual but also in his eyes emotional opposition, Turing made extensive use of irony and, specifically, satire, or irony with a point. Sensing that he was not properly listened to, some of his communications were assembled to shock. Having borrowed myself a perspective developed to interpret the irony of David Hume, I have argued that Turing's irony shall neither be understood always literally nor dismissed as parody or plain mockery. He was far from being a hermetic writer. I claimed that the interpretation of Turing's irony is a tractable problem.

Style of reasoning (§1.4). I pointed out that Turing's style of reasoning has been articulated in a most elegant way by Max Newman. In order to study "the extent and the limitations of mechanistic explanations of nature," Turing's "way of tackling the problem was not by philosophical discussion of general principles, but by mathematical proof of certain limited results" (1955, p. 256). (And this responds to the second interpretation suggested by Turing's contenders, whether his views were significantly mixed with some sort of wishful thinking, cf. the end of §1.2.) I also relied on the Crombie-Hacking enumeration of styles of scientific reasoning to describe Turing's in terms of the history and the philosophy of science. He was a mathematician, but also an empiricist. He proceeded by observing nature and establishing conjectures that were testable both analytically and empirically. He made use of analogies and constructed hypothetical analogical models. He was a master of thought experiment. He posed a serious question to the philosophy of mind, and yet, as a philosopher, Turing thought like a scientist.

Ambition (§1.5). Turing implied that machines shall not be condemned to be slaves forever. He posed a fact-value demarcation. He distinguished the problem of studying the possibility of us humans being superseded by machines or other animal species in intellectual power, on the one hand, and whether or not it would be agreeable to someone or to mankind, on the other hand. He admitted "that its realization would be very disagreeable" (2004 [1948], p 410), but for him what was at stake did not depend on how likable it was. (And this responds to the first interpretation question suggested by Turing's contenders, whether he considered dangerous social and ethical implications of his views, cf. the end of §1.2.) He challenged an anthropocentric attitude towards machines and animals. He referred to Samuel Butler's *Erewhon* and presented a shocking scenario in which the machines eventually supersede us humans in intellectual power and take control over us. I hope to have shown that Turing was responding satirically to a 1949 provocation by the BMJ editorial. And yet, for Turing, the possibility of having machines outstripping our powers and taking control at some point in the future was true. While he thought so and in fact wanted to build intelligent machines, he did not actually want

the machines to take control. This was the limiting case of his scientific argument and also of his social critique against species and gender chauvinism. Because intelligent machines were a refutation of biases of species and gender, his future society pervaded by them may be seen as a dystopia but also as a utopia. He rather took superintelligent machines as something serious that needs to be studied in detail, both analytically and empirically. He thought that such possibility was not astronomically remote, but remote enough so that the benefits for human enlightenment would pay off in the meantime. Contributing to knowledge about the human mindbrain along the lines of the modern tradition of mechanistic natural philosophy was his most primary aim. But Turing was no court natural philosopher. Because he also found himself turned to be a social critic, Percy Shelley's enlightened Prometheus suits him very well indeed. As for negative interpretation boundaries, I have shown suggestive evidence that Turing's ambitions can be dissociated from the image of Mary Shelley's Dr. Frankenstein. He thought that fantastic projects, *e.g.*, making a machine resembling the human body, was futile and bound to give results whose quality would be unpleasant like artificial flowers. The evidence gathered is also unfavorable for associating Turing with contemporary transhumanist movements in general. Turing was a scientist who wanted to study the status of machine intelligence as the other face of the status of human intelligence. Meeting with a conservative reaction, he made a subtle social critique.

Plea (§1.6). Turing pleaded for a fair play for the machines. Systematically in the rebuttal of five out of the nine lines of attack that he considered against his views on machine thinking, Turing pushed back by shifting the burden of proof towards human thinking and the human mind. In following this specific argumentation tactics, he became even more socially inconvenient. He did that in a negative sense to challenge his contenders' views of ourselves because he thought that some reactions were "purely emotional" (2004 [1948], p. 411). But also he did it in a positive sense because he considered that "the idea of 'intelligence' is itself emotional rather than mathematical" (*Ibid.*). For him, arguing for the possibility of machine thinking might require reviewing what is meant by (human) thinking. This could lead in turn to reviewing the ontological status of human, and specially, it could challenge a certain notion of free will. Thus, in spite of having fought bravely for his country against Nazi Germany in World War II, he has been suggested to promote ideas that were thought to lend themselves to Nazism. In view of informing the historical context of the reception of Turing's views on intelligent machines, I have pointed out that such specific form of *ad hominem* argument is known as *reductio ad Hitlerum*. I have also highlighted the social and cultural background against which Turing presented his bold views on intelligent machines, namely, Cold War Britain. Back then, any kind of "non-conformism" could be associated to lack of patriotism and possible involvement with Communism by Cold War hysteria. While this background shall not be crucial for the assessment of the reception of Turing's test, it is definitely something that shall be kept in mind.

Fate (§1.7). Turing was a revolutionary thinker who suited Galileo's eagle-starlings metaphor about what good philosophers are like. He faced opposition in Postwar Britain because of his views on machine thinking, which promised to have deep consequences to our view of

ourselves. Having been sent to trial and punished for being gay, we know from his *c.* 1952 letter to a friend that he connected that specific persecution with the facile rejection of his views on intelligent machines. The *ad hominem* attacks that he had been receiving early on would now become overwhelming and unmanageable. The real reason for his death remains unknown. We can say that he lived and died in the pursuit of his beliefs.

1.9 Epilogue

Recently Mario Livio produced a stunning development for an old mystery (2020). He has been able to discredit the story of a painting, *Galileo in prison*, previously dated to *c.* 1645 (only a couple years after Galileo's death), which would have been the earliest printed mention of a motto attributed to Galileo in connection with his trial. The story had deceived generations of renowned Galileo scholars such as Antonio Favaro, John J. Fahie, and Stillman Drake. The portrait has been attributed to Spanish painter Bartolomé Esteban Murillo (Cf. FAHIE, 1929, plate XVI, with presentation in p. 72-5). It would have had an extra frame that hid its heretical contents and saved it from destruction while Galileo's *Dialogue* was still in the *Index Librorum Prohibitorum*. It would have been hidden away in the caves of the Société Générate de Belgique in Brussels in time to escape the ravages of World War I. Drawn on the wall of Galileo's prison in the portrait was a scheme of the solar system. Written on the wall, just beneath Jupiter and its satellites, was the legendary *E pur si muove!* ("And yet it moves!", meaning the Earth). Contrary to Fahie's absolutely beautiful gathering of the evidence that was then available to him in (1929) it turns out that the portrait is most likely original from Flemish painter R.-E. van Maldeghem in 1837, just a couple years after the *Dialogue* has been omitted from the *Index* (Cf. FINOCCHIARO, 2005, p. 193). The phrase attributed to Galileo is likely apocryphal, indeed. It lives as a symbol.

Mutatis mutandis, should Turing had lived to fully recover his social legitimacy and confidence after trial and public exposure for being gay, I can well imagine him replacing his hopeless 1952 syllogism that ended "machines do not think" by phrase "And yet they think!" Arguably, Turing's question is hardly less important than Galileo's. As with the Renaissance thinker, one may hope that, in the end, truth will conquer. But was Turing right? We shall continue this study and examination towards hopefully a sensible appraisal of his hypothesis and its famous test.

1.10 Chapter acknowledgements

I am indebted to Pedro Bravo for pointing me to the specific form of *ad hominem* argument called *reduction ad Hitlerum*, and for the suggestion of John V. Price's *The ironic Hume* after having heard me to talk about Turing's irony. I owe to Prof. Pablo Mariconda my appreciation of Galileo as an intrepid thinker in general, and his *Assayer*'s eagle-starlings image in particular. Possible mistakes in my elaborations out of these materials are my sole responsibility. I also

thank very much the Turing biographers cited herein for contributing with their valuable work in collecting primary and secondary sources, without which this research could not have been done.

2 Turing's existential hypothesis on thinking machines

Can machines think? As known, in his seminal 1950 paper Turing avoided to address the question directly. He rather handled it, as I shall argue (§3), *from within* his iconic imitation-game thought experiment. But in other moments of his dialogical years (1949-1952; §A.4), Turing did address his original question directly. In this chapter I shall identify and study key passages from Turing sources to arrive at a philosophical interpretation of Turing's hypothesis that "machines can think." One such passage runs as follows.

The discussion was going on in the seminar "the mind and the computing machine," on the 27th of October in 1949 at the University of Manchester, Department of Philosophy. At some point, according to the minute notes that survived to us, Michael Polanyi made a second reference to his notion of "semantic function," and Turing gave a most spontaneous reply:

> POLANYI: interprets this [an observation by Turing on consciousness and machines] as suggestion that the semantic function can ultimately be specified; whereas in point of fact a machine is fully specifiable, while a mind is not.
>
> TURING: replies that the mind is only said to be unspecifiable because it has *not yet been* specified; but it is a fact that it would be impossible to find the programme inserted into quite a simple machine – and we are in the same position as regards the brain. The conclusion that the mind is unspecifiable does not follow.
>
> POLANYI: says that this should mean that you cannot decide logical problems by empirical methods. [...] (TURING et al., 2005 [1949], no emphasis added)

Here lied, I think, a crucial exchange. Polanyi was saying it to the face of the scientist who led a team of cryptographers into breaking the logic behind the German machine-enciphered messages in the Second World War, and they did it by observing the machine's behavior alone. To say it differently, they were able to approximate close enough for their purpose the inner logics of machines (their embedded programs as instantiated with input settings and data) by using nothing but empirical methods. And yet the program (if any) inserted into the human brain would be, of course, a much more challenging task, if not the hardest one ever, indeed. Turing's scientific project was of vaulting ambition. Interlocutors were outraged. The conversation resumed:

> POLANYI: [...] The terms by which we specify the operations of the mind are such that they cannot be said to have specified the mind. The specification of the mind implies the presence of unspecified and pro-tanto unspecifiable elements.
>
> TURING: feels that this means that *my* mind as *I know it* cannot be compared to a machine.
>
> POLANYI: says that acceptance as a person implies the acceptance of unspecified functions.

> ....?: re-raises the point regarding the undiscoverability of a programme inserted into machines. Could this be clarified?
> (TURING et al., 2005 [1949], no emphasis added)

Turing's point about the intractability of discovering a program was that there was no way to find out or reverse engineer a computer program by observing its outer behavior alone. By "mind," Polanyi seems to have meant an introspective, subjectivist or personified mind. Turing seems in turn to have meant a generic, objective or naturalized mind. And would the logic or program (if any) corresponding to the natural mind, as (if) inserted into the natural brain, be discoverable? As we see from the passage, Turing thought not. But his rationale seems less related to mysteries surrounding the mindbrain and more related to its natural complexity. Accordingly, based on in this exchange with Polanyi, Turing does not seem to have set out to discover a supposedly exact human mindbrain program. Although specifiable, so he thought, it entails an empirical problem that would be impossible to solve. So, one may ask, what did Turing target exactly?

2.1 Problem and chapter structure

Perhaps the most common philosophical reading of Turing has been that his views of thinking or intelligence are to be classed, according to the jargon in different areas, as positivism, phenomenalism, or operationalism (philosophy and physics), or as behaviorism (psychology). This reception of Turing's views is usually based on a reading of his 1950 paper, which also for historical reasons (as other sources have not been widely available until recently) has been the canonical source for Turing's views on mind and machine. In fact, lying in the forefront of Turing's 1950 text is his proposal of the imitation game or test for machine intelligence. Most supporters of Turing's proposal as a decisive test, as we shall see (§3.2), have contended that Turing did *not* have an operational or behaviorist view of thinking or intelligence because he did not offer any such (operational) definition. This view, they argue, would require one to find both a necessary and a sufficient condition for thinking in Turing's 1950 proposal. But Turing would have actually offered only the latter. And yet, feeling that Turing engaged in behaviorism all the same, critics managed to construe a notion of sufficiency behaviorism in his proposal anyway.

Now, on the one hand, reading Turing's 1950 proposal as operationalist or behavorist seems to imply the view that he did not have or refer to any ontology of thinking or intelligence at all, for any such talk, according to the doctrine of behaviorism, would be "meaningless." On the other hand, reading it as *not* operationalist or behavorist and yet accepting it as a test and/or as an empirical basis for thinking or intelligence seems to necessarily imply the view that Turing's 1950 proposal was part of a methodology and/or epistemology that was packed with some ontology in the backside. We thus arrive at what I shall call:

The problem of Turing's hypothesis. What did Turing mean by his hypothesis that "machines can think" exactly, and what was his philosophical attitude towards it from a philosophy

of science point of view? Did Turing have an ontology of thinking, and/or of the mind?

For some early clarification of terminology relative to my formulation of the problem, I shall mostly follow Turing's own usages. There is no clear sign of a mind-brain distinction in his writings, lectures, letters etc.. In representative passages (*e.g.*, in his 1950 "skin of an onion" image, p. 454), he referred to "the functions of the mind or the brain." So mentions of "mind" in other passages by Turing are likely to be deflationary uses of the term as he engaged in discussion with the terminology of others. While I am promptly suggesting to read Turing from this sort of neutral-monism interpretation at this point, the contention will present itself in more depth as we go further in the studies of this chapter. Note also that my focus is on his philosophy of science of the mind(brain). I shall thus structure my discussion along these lines. Jargon from the philosophy of mind may come to the forefront in specific moments of the development of my argument — specially in connection with Turing's ontology and attitude towards his hypothesis. A focused discussion with the philosophy of mind literature, however, is beyond my scope in this dissertation. Turing does not seem either to have drawn a distinction between "thinking" and "intelligence." For him, I shall assume, an intelligent entity is one that can be said to think, and a thinking entity is one that is to be attributed the property of having intelligence.

I shall give now short answers to the questions formulated in the problem of this chapter, as I simultaneously introduce the structure of my argument and its development throughout this chapter. To develop my interpretation of what Turing meant exactly by his hypothesis on thinking machines, I shall propose a clear-cut distinction between *Turing's epistemology and his ontology*. By these categories I mean, as usual, a theory of knowledge and a theory of being. So, in relation to Turing's hypothesis, if we look at it from an epistemological point of view, we are trying to understand Turing's proposed view on how to justify the attribution of thinking or intelligence to an entity as a knowledge claim. If we look at it in turn from an ontological point of view, we are trying to understand Turing's proposed view on what thinking or intelligence (really) is. I am not aware of any previous explicit reference to this distinction relative to Turing's test and hypothesis. My hope is that it will be fruitful towards an understanding of Turing's views and their reception in the secondary literature. Accordingly, I will start with a quick review of representative readings of Turing with an emphasis on my proposed distinction (§2.2).

It will come next the first step of my argument (§§2.3, 2.4). I will try to show that, although Turing did not explicitly distinguish what he meant in terms of epistemological or ontological stances, he did have two quite different concepts of thinking or intelligence. Turing's epistemology of thinking is based on causing wonder on an external observer. It is subjectivist and emotional, yet observable. Wonder is the central concept of Turing's proposed evidential basis. It is upon this basis that machine intelligence was to be put to test. Turing's ontology of thinking in turn is based on learning from experience like a brain. It is objectivist and mathematical, yet only indirectly observable. Learning is the central concept of Turing's notion of machine thinking or intelligence. The strength of Turing's view lies in the connection of these two foundational

concepts (wonder and learning) which, as I shall emphasize, are strongly coupled to one another by design. Together they constitute, I hold, Turing's philosophy of science of the mindbrain. I will quote extensively from primary sources to support this view on the epistemological and ontological foundations that Turing proposed towards building and testing thinking machines.

As a second step of my proposed interpretation of Turing's hypothesis, I shall examine his philosophical attitude towards it (§2.5). This amounts to exploring his statements, their tone, his expectations, their targets and so on. This I shall also do from a philosophy of science point view, in connection with the developments of the previous sections. I will cast Turing's attitude as realist, meaning that he proceeded by conjecturing a natural mechanical mindbrain to really exist in the human head and be amenable for empirical study. At this point, it shall become apparent that Turing's target, in my view, was less a philosophy and more a physics of the mind. I will try to show that Turing set out to specify (or to find) a model or digital replica of mechanical mindbrain to mirror the one he conjectured to exist as part of the real human mindbrain. Turing's notion of mechanical mindbrain is a hypothetical construct, but it is strongly tied up to the empirical or evidential basis he proposed for machine intelligence. All of this, of course, is implied in his test. Turing's model of the human mechanical mindbrain would be able to approximate very closely the real one such that their performances in a wide class of most impressive intellectual tasks would be indistinguishable. For him, *n.b.*, the real mechanical mindbrain — which is "a continuous machine" — *could not* be literally a digital computer — which is "a discrete-state machine" — and thus cannot be exactly replicated but only mimicked. Note that it is precisely because the ontological status of the mechanical mindbrain digital replica (or machine thinking) in Turing's view is different than actual identity with the real mechanical mindbrain (or human thinking) that the correct reference to it according to Turing, I hold, is by a notion of $thinking_2$, rather than by $thinking_1$, as suggested very early on in the Introduction.

Equipped with the textual and conceptual interpretive basis built in the sections covered so far, I shall proceed to push it further in depth in connection with Herbert Feigl's 1950 classical account of existential hypotheses in the empirical sciences (§2.6). Feigl offered a precise notion of this specific class of hypotheses. They refer to theoretical and only indirectly observable constructs which are posited in laws that are themselves only indirectly testable. He then surveyed nine possible interpretations of such hypotheses, informed by the history of the empirical sciences and their philosophical study. My hope is that Feigl's terminology and concepts will prove worth it for the interpretation of Turing's views and their reception. I shall then discuss how existing interpretations of Turing's epistemology and ontology as reviewed before may fit into Feigl's interpretation classes (§2.7). Given all the textual and the interpretive basis developed, I will argue that Turing's hypothesis on the mechanical mindbrain and his philosophical attitude towards it can be best associated with Feigl's specific empirical realism (§2.8). Turing postulated, I shall hold, a machine-brain identity when it comes to intellectual power. Accordingly, he saw a correspondence of empirical properties that he postulated between digital computers and the human mindbrain. These two entities, Turing conjectured, shared some

common ontological status.

Altogether, I hope to contribute to show that Turing addressed the human mindbrain from a natural science point of view. And yet that does not mean that his views did not have a social epistemology component. I shall conclude this study of Turing's hypothesis on the existence of thinking machines with a few remarks about his views on the subjectivity of machines (§2.9).

2.2 Received views of Turing's epistemology and ontology

As I have suggested, the operationalist or behaviorist interpretations of Turing's 1950 proposal found no ontology at all in it. According to this reading, Turing would have actually reduced thinking or intelligence to verbal behavior, and any talk of hypothetical construct of mind and its internal states, either in the human or in the machine, would be misleading or meaningless. Usually these interpretations capitalize on Turing's phrasing (1950, p. 442): "[t]he original question, 'Can machines think?' I believe to be too meaningless to deserve discussion." (We shall examine this sentence in §3.3.) A related interpretation that sometimes is made of Turing's 1950 text is that at a first glance he seems to be offering an operationalist definition of intelligence, but later on — notably in his discussion of the nine objections, in particular when rebutting to the arguments "from consciousness" and "from various disabilities" — he would have slipped into committing to an ontological position in the philosophy of mind. But this charge violates the elementary exegetical principle of assuming coherence and intelligibility in an author's text. Turing could not have been at the same time offering an operational definition of thinking and committing to an ontology of mind, as this charge would imply viewing his text as incoherent.

Another possibility of interpretation that was available for Turing relative to his machine-brain analogy is instrumentalism. This is worth mentioning because it was pushed to him by his contender, the neurosurgeon Geoffrey Jefferson. In the ending of the 1952 BBC radio roundtable (§A.4.6), Jefferson pushed to Newman and Turing an instrumentalist view of the new electronic computing machines, as opposed to, say, Turing's view of the machine as an electronic brain:

> *Jefferson*: [...] You have the great advantage of knowing how your machine was made. We only know that we have in the human nervous system a concern compact in size and in its way perfect for its job. We know a great deal about its microscopical structure and its connections. If fact, we know everything except how these myriads of cells allow us to think. But, Newman, before we say 'not only does this machine think but also here in this machine we have an exact counterpart of the wiring and circuits of human nervous systems,' I ought to ask whether machines have been built or could be built which are as it were anatomically different, and yet produce the same work.
> *Newman*: The logical plan of all of them is rather similar, but certainly their anatomy, and I suppose you could say their physiology, varies a lot.
> *Jefferson*: Yes, that's what I imagined — we cannot then assume that any one of these electronic machines is a replica of part of a man's brain even though the result of its actions has to be conceded as thought. The real value of the machine to you is its end results, its performance, rather than that its plan reveals to us a model of our brains and nerves. Its usefulness lies in the fact that electricity

> travels along wires 2 or 3 million times faster than nerve impulses pass along nerves. [...] But that old slow coach, man, is the one with the ideas — or so I think. It would be fun some day, Turing, to listen to a discussion, say on the Fourth Programme, between two machines on why human beings think that they think! (TURING et al., 2004 [1952], p. 505-6)

Contrary to Jefferson's dragging, Turing had never endorsed an instrumentalist view of modern computing. Quite the opposite. His contention, not only with Jefferson, but also with, for instance, Douglas Hartree (another computer pioneer), was to stand against this from the beginning.

There have been interpretations that either explicitly or implicitly rejected the view of Turing's 1950 proposal as naive operationalism. Those interpretations have seen it as the proposal of a test and/or an empirical basis for thinking or intelligence but they varied in their answers to the problem of his philosophy of science of the mind. There has been — and it is tentatively and unpretentiously that I name them as follows — an inductive-inference view, a social-epistemology view, and more recently, an epistemic-ontological view and a normative-constructionist view. Let me take a moment to very briefly review them in chronological order before I introduce my argument and the chapter structure. They all have historical and intellectual significance that go beyond my review here, which is tailored to my focus.

According to James Moor (1976), "the real value of the imitation game lies not in treating it as the basis for an operational definition but in considering it as a potential source of good inductive evidence for the hypothesis that machines think" (p. 249). Moor meant "our ordinary concept of thinking," which he took to be "to process information in ways which involve recognition, imagination, evaluation and decision" (p. 250). Moor suggested a variant of the test (for example, with no gender question at all) based on a notion of inductive inference over probabilities: "On the average after n minutes or m questions is an interrogator's *probability* of correctly identifying which respondent is a machine significantly greater than 50 percent?" (p. 249, emphasis added). Moor also suggested that thinking would, say, be reified in the computer as thought: "the Turing test is a significant test for computer thought if it is interpreted inductively" (p. 256). Thinking or thought is thus an unobservable property or entity whose presence is made observable in the test.

Judith Genova produced in (1994) the social-epistemology interpretation of Turing's test. She understood that Turing corresponded thinking to the imitation of thinking, taking them as equivalent. Let me quote:

> Unlike dialectic, which is intent on giving everything its proper place, computing is uninterested in the proper and thus holds no mirror to nature. Build it anyway you like as long as it works, i.e. as long as it emulates the behavior or the properties of the original. So-called natural differences are irrelevant to this process. The differences between apples and oranges, for example, or those between men and women, disappear in the realm of numbers. [...]
>
> [...] Good simulation collapses the distinction between the real and the simulated. However, the only reason the machine can become human is because humans

> are already machines. To be a symbolic animal is to be the animal that can transform itself into what it is not. (GENOVA, 1994, p. 320)

This comes close, I think, to the functionalist interpretation of Turing in the philosophy of mind. Genova offered the remarkable insight that, for Turing as a homossexual, it must have been natural to consider that physical or biological constitution shall not determine function or behavior. She observed: "Indeed, the fact that gender is a matter of knowledge suggests that both thinking and being for Turing are discursive, cultural phenomena, not biological ones." Rather, she concluded, "to put the point more succinctly for Turing, biology is open to thought's manipulation" (p. 315). Regarding computing and simulation, however, Genova's interpretation that emulation or good simulation "collapses the distinction between the real and the simulated" is not the only possible interpretation of computer simulation. In the modern computational sciences good simulation is rather often viewed as a tentatively posited mirror to nature, distinct from nature itself. The mirrored image is constructed to serve as a cost-effective basis for explanation and prediction, and is successively corrected.

Jack Copeland gave in (2000b) a concise view of Turing's philosophy of science of the mind, cast as a principle:

> *Turing's Principle*: A machine that by means of calculation imitates — or, better, 'emulates,' for Turing is concerned with faithful imitation [Copeland's note: Elsewhere Turing uses the verb 'simulate': 'My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely'] — the intellectual behaviour of a human brain can itself appropriately be described as a brain, or as thinking. (Only the intellectual behaviour of the brain need be emulated: 'we are not interested in the fact that the brain has the consistency of cold porridge'.) (COPELAND, 2000b, p. 529-30)

In fact, whatever the machine does, it must be a form of calculation or mechanical operation. And yet I do think that Turing was specific about *how* the machine is supposed to imitate (by mechanical operation) the brain so that it qualifies as thinking. Copeland presented historical evidence against the view that Turing would have offered an operational definition of thinking or intelligence (p. 522-3). And he considered Turing's principle to be a core part of what he called "the foundation of the Turing test" (p. 529). He quoted at length from Turing's May 1951 BBC radio lecture "Can digital computers think?" (§A.4.5), which I take to be, indeed, the central source for the interpretation of Turing's philosophy of science of the mind. If for nothing else, it is a primary source and the one in which Turing seems to have talked most openly and free of irony. I shall soon collect key passages from it (§§§2.3, 2.4, 2.5). Copeland combined (1) Turing's principle above with "(2) the claim that the method of question and answer provides a suitable means for determining whether or not a machine is able to emulate human intellectual behaviour;" and furthermore with "(3) [...] the imitation game," which "in its specific provisions, is suitable for this purpose" (p. 530). So for him, I interpret, the conversational question-answering imitation test defines Turing's epistemology to evaluate whether or not a machine qualifies as being able

to imitate a brain, or to think. But what does the status of such an emulation competence really mean? In (2000a), Copeland interpreted Turing to have suggested that "it is 'not altogether unreasonable' to describe a machine that 'imitate[s] a brain' *as itself being a brain*" (p. 31, emphasis is mine). For context, Copeland added next, quoting from the 1952 BBC roundtable with Jefferson and others: "(As is well known, Turing advocates imitation as the basis of a test that '[y]ou might call [...] a test to see whether the machine thinks'." Now, shall we take this to mean that Turing proposed imitation in general, as we have just seen from Judith Genova, as an ontology where "the distinction between the real and the simulated" is collapsed? I think not, for later in (2013), Copeland suggested relations between Turing's thought and a physics of the mind. He wrote: "Turing [...] seems to have believed that the mind is a partially random machine" (p. 657). And added: "[w]e have the word of one of Turing's closest associated, Newman, that Turing 'had a deep-seated conviction that the real brain has a "roulette wheel" somewhere in it.' (The source for Newman's testimony was an interview to Christopher Evans that composes *The pioneers of computing: an oral history of computing*, available in the Science Museum in London.)

According to Darren Abramson (2011), Turing's 1950 proposal can be best understood as addressing the same problem of Descartes's. In his words: "Turing at least conceived of his own test as fulfilling just the epistemic purpose that Descartes' fulfilled for him" (p. 550). As Descartes proposed his "language test" as an empirical basis to provide (unfavorable) evidence of "the presence of some property that is necessary for mind" in machines (and in non-rational animals), Turing would have proposed his own to provide (favorable) evidence of the same kind. This property, Abramson argued early since (2008), is Turing's "commitment to a necessary condition for thought." And because it is an non-behavioral property, Turing cannot have been a behaviorist. Abramson called this inner property "the epistemic-limitation condition," and pointed that Turing would have revealed it "in his response to 'Lady Lovelace's objection"' (p. 546). The condition is a representation, Abramson argued, of the relationship between the machine's exhibition of intelligent behavior and the knowledge of the person who designed it. If the actual behavior of a machine goes beyond what its designer can anticipate or predict — that is, if it goes beyond the dispositional behavior of the machine as seen by its designer —, then machines can be said to think. Abramson's view stemmed from having studied what he took to be the historical influence of Descartes on Turing. He considered that "Descartes' views at least helped crystallize Turing's own conception of the Turing test, and at most presented him with the idea *in toto*" (p. 549, no emphasis added). Abramson took to be an "irony" that "Turing, an apparent materialist about mind, and Descartes, a dualist, agree on how we can determine that machines do or don't have minds" (p. 548). But if Turing was a materialist about mind, would the latter even exist at all in any sense comparable to Descartes's view? In fact, Abramson did not suggest any ontological basis for the mind from Turing's point of view. The latter, as I understand from Abramson, seems to be nothing but a reification of Turing's epistemic-limitation condition.

In (2013) Diane Proudfoot considered that Turing did not distinguish thinking and

intelligence as two different notions, and that his concept of thinking or intelligence is emotional rather than mathematical. Accordingly, she produced a full-fledged new interpretation of the Turing test. For her, Turing's notion of thinking or intelligence can be best understood in terms of response-dependency theory, and the test is thus nothing but the specification of a particular response-theoretic scheme for that concept. This means essentially that a machine can be said to think or to be intelligent if it is thus construed by normal observers under normal conditions of observation. There is an analogy with inter-subjective concepts, for example, a color, which can be perceived similarly by people who are not colorblind in adequate lighting conditions. And that level of description does not preclude another — and related — concept of color, say, one that has a physics. Likewise, Proudfoot's view would in principle not preclude a physics of thinking or intelligence. In my view, as we shall see (§2.3), Turing seems to have explicitly suggested this complementary principle, so to speak. He pointed out that the attribution of intelligence "is determined as much by our own state of mind and training as by the properties of the object under consideration" (2004 [1948], p. 431). But Proudfoot seems to have thought otherwise:

> It is not a test of intelligence that we need, some critics say, it is a (computational) *theory* of intelligence. This, however, assumes that such a theory is possible — and if intelligence is an emotional concept then such a theory is not possible. (PROUDFOOT, 2017a, p. 295, no emphasis added)

Turing's view, in Proudfoot's interpretation as it seems, intelligence is to be ruled normatively and decided inter-subjectively but never objectively. (By an inter-subjective decision here, I mean one depending on subjective human judges, even if they rely on a previously agreed upon scheme of rules. By an objective decision, I mean one depending solely on the properties of the object under consideration as measured by a scientific instrument. Of course, instruments are designed by embedding concepts and standards that are previously defined by us, but note that once they are embedded into the device, the act of measurement is made steady and objective in principle.) In fact, for Proudfoot (2017b), Turing's test measures the observers, not the machine. For this reason, she argued, Turing could not have offered a behaviorist conception of intelligence.

Abramson and Proudfoot's interpretations of Turing's concept of thinking or intelligence contributed for an understanding of the evidential basis that Turing proposed for machine intelligence. Now, for a comparison of their views, one difference, I think, lies in how to control for subjectivity in the test's decision scheme. Proudfoot suggests to do it by means of a predefined scheme of normal observers under normal conditions of observation that, she argues, has already been indicated by Turing in his 1950 text. Abramson in turn requires that the observer has to be the designer of the machine for, he argues, only she could know for sure whether or not the epistemic-limitation condition has been satisfied, that is, whether or not the machine has brought about something really new. This difference, from a general outlook on the highly heterogeneous manifold of interpretations of Turing's test, is not very significant. It can be seen as a specific difference in how to implement the epistemological criterion of the test. For a side

note, they both tried explicitly, in their own ways, to refute the view of Turing as a behaviorist. In effect, their views of Turing's ontology appear to be similar, up to Abramson's reification of the epistemic-limitation condition into "mind." According to their interpretations, intelligence, for Turing, would not even exist if there was not for an external observer's mind to take notice. They both seem to have blurred a distinction between Turing's epistemology and his ontology (if any).

In the next sections I will present my own view on all this. I shall work out a clear-cut distinction between Turing's epistemology and his ontology, and suggest that neglecting Turing's ontological concept of thinking may have contributed to the (distorted) view that Turing's research program on machine intelligence was to build a machine simply to deceive.

2.3 Wonder: Turing's epistemology of thinking

However learned a machine can be, there may be emotional barriers for one to attribute thinking or intelligence to it. In his 1948 *Intelligent machinery* report, the two first objections to the possibility of machine intelligence that Turing outlined were (a) and (b), or what Turing called in 1950 the "heads in the sand" and the "theological" objections (cf. Figure 1). He wrote:

> The objections (a) and (b), being purely emotional, do not really need to be refuted. If one feels it necessary to refute them there is little to be said that could hope to prevail, though the actual production of the machines would probably have some effect. In so far then as we are influenced by such arguments we are bound to be left feeling rather uneasy about the whole project, at any rate for the present. These arguments cannot be wholly ignored, because the idea of 'intelligence' is itself emotional rather than mathematical.
> (TURING, 2004 [1948], p. 411)

This was the specific context of discussion in the beginning of Turing's 1948 text, where he stated that intelligence is an emotional concept. Then he spent the core part of his argument by developing how machine learning could be achieved through *mathematical* modeling. It was only further on in his text, in its concluding section §13 entitled "[i]ntelligence as an emotional concept," that Turing returned to this view of intelligence. He thus wrote:

> The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration. If we are able to explain and predict its behaviour or if there seems to be little underlying plan, we have little temptation to imagine intelligence. With the same object therefore it is possible that one man would consider it as intelligent and another would not; the second man would have found out the rules of its behaviour. (TURING, 2004 [1948], p. 431)

In this passage Turing considered *the attribution of intelligence* to a given entity by an external observer to depend on two elements, one subjective ("our own state of mind and training" as observers) and the other objective ("the properties of the object under consideration"). These two elements were, for Turing, equally important (*n.b.*, in the passage, "as much"). Given the

structure of Turing's 1948 text, his goal in this particular passage seems to have been to stress that the objective component — to which he dedicated most and the core of his text — is not enough for a machine to be granted the property of thinking or intelligence.

Turing would return to his view of intelligence as an emotional concept in his 1950 paper, when discussing the fifth objection, "[a]rguments from various disabilities." (This passage, to my knowledge, has not been identified in this connection in the secondary literature.) Most interestingly, in this occasion, Turing related it to Jefferson and the argument from consciousness:

> The criticisms that we are considering here are often disguised forms of the argument from consciousness. Usually if one maintains that a machine can do one of these things, and describes the kind of method that the machine could use, one will not make much of an impression. It is thought that the method (whatever it may be, for it must be mechanical) is really rather base. Compare the parenthesis in Jefferson's statement quoted on p. [445-6, namely, "(and not merely artificially signal, an easy contrivance)"]. (TURING, 1950, p. 449-50)

He would insist on this point in the 1952 BBC roundtable with Geoffrey Jefferson, Max Newman and Richard Braithwaite. Eventually in their discussion, Braithwaite asked and Turing replied:

> *Braithwaite*: But could a machine really do this [learn by analogy]? How would it do it?
> *Turing*: I've certainly left a great deal to the imagination. If I had given a longer explanation I might have made it seem more certain that what I was describing was feasible, but you would probably feel rather uneasy about it all, and you'd probably exclaim impatiently, 'Well, yes, I see that a machine could do all that, but I wouldn't call it thinking.' As soon as one can see the cause and effect working themselves out in the brain, one regards it as not being thinking, but a sort of unimaginative donkey-work. From this point of view one might be tempted to define thinking as consisting of 'those mental processes that we don't understand'. If this is right then to make a thinking machine is to make one which does interesting things without our really understanding quite how it is done. (TURING et al., 2004 [1952], p. 500)

Now, there are three aspects that I want to draw attention to in this passage. The first, on the one hand, is that Turing thereby consolidated this non-obvious view of intelligence as an emotional concept, which he had expressed already in 1948 and in 1950. Second, he also suggested, on the other hand, that this is but one "point of view" and it may not be necessarily right — in the sense, I guess, of not being itself enough to account for intelligence. For a central shortcoming, it can be taken to the limit to imply that intelligence is whatever mental processes that we do not understand, which would make it trivial. (Even a psychiatric disorder may be a mental process that we do not understand. But Turing seems to have been well aware of the limitations of such elusive view of intelligence.) Third, the passage makes it clear that Turing's view of intelligence as an emotional concept is negative in a certain sense. For Turing, intelligence is only construed or "imagined" in a dialectic between observed and observer(s). It involves the complement of what the designator(s) of the property of intelligence know(s) or can anticipate at the time of designation. Indeed, Turing's concept of intelligence as emotional seems to have been conceived

to prevent or block *confirmation bias* towards the conservative position that machines never thought, can't think and will never think. This was well remarked by Max Newman after Turing also in the BBC roundtable against one of the demands of Jefferson's:

> *Jefferson*: [...] I don't see how a machine could as it were say 'Now Professor Newman or Mr. Turing, I don't like this programme at all that you've just put into me, in fact I'm not going to have anything to do with it.'
> *Newman*: One difficulty about answering that is one that Turing has already mentioned. If someone says, 'Could a machine do this, e.g. could it say "I don't like the programme you have just put into me", and a programme for doing that very thing is duly produced, it is apt to have an artificial and ad hoc air, and appear to be more of a trick than a serious answer to the question. It is like those passages in the Bible, which worried me as a small boy, that say that such and such was done 'that the prophecy might be fulfilled which says' so and so. This always seemed to me a most unfair way of making sure that the prophecy came true. If I answer your question, Jefferson, by making a routine which simply caused the machine to say just the words 'Newman and Turing, I don't like your programme', you would certainly feel this was a rather childish trick, and not the answer to what you really wanted to know. But yet it's hard to pin down what you want. (TURING et al., 2004 [1952], p. 501-2)

So Turing (and Newman too) observed that demands such as Jefferson's would pose *ad hoc* and biased barriers for the discussion. In fact, it is to the extent that we humans are emotional — and this, I think, can hardly be questioned — that the capability to cause wonder, or to take us by surprise, or, in the context of Turing's passages dressed in irony, to deceive, is all that is required for the observation and the designation of intelligence. This capability is, therefore, required for setting up a directly observable (external) empirical basis for intelligence. The two objections ("a" and "b") that Turing classed as "emotional" as of 1948 were formulated given his perception of the conventional wisdom of the time. If today there may be bias (among intellectuals) against acknowledging some performance or feat by a machine as "thinking," one might wonder how much there was as of the turn from the late 1940's to the early 1950's when computing machines had their valves exposed and barely a screen. Turing seems to have realized how important — socially, culturally, psychologically — the concept of intelligence was and that, accordingly, there was a need to prevent or block the biases involved in attributing it. Altogether, I hold, Turing's view of intelligence as an emotional concept is but one component of his dual concept of thinking or intelligence. Equipped with this mathematically-empty notion, neither Turing nor anyone could get anywhere near building an intelligent machine. However, since that project was the very reason why Turing eventually found himself debating whether machines can think, the interpretation of his view of intelligence as an emotional concept *as exhaustive of* his concept of intelligence does not seem to make sense at all. It would rather imply that Turing's research program would better have been to build a machine not to learn in general, but simply to deceive. This, as we shall see later (§3.2), has been a charge made on his test in the related literature, and I hope overall to show that it is false. Turing himself rejected any resort to "easy contrivances" and, in fact, called it just "cheating," as we shall see shortly (§2.4).

Turing's reduction of the appearance of free will into intelligence

Turing also expressed his view of intelligence as an emotional concept in the multiple versions of his rebuttal to Lady Lovelace's objection to machine intelligence. He suggested that the machine must be able to "take us by surprise." The mostly developed version of this appeared near the end of his May 1951 BBC radio lecture (§A.4.5). In this occasion, Turing identified an interaction between the problem of making a machine to display intelligence or to think and the problem of making it "appear as if it had free will" (2004 [1951], p. 484). He said: "[t]o behave like a brain seems to involve free will, but the behavior of a digital computer, when it has been programmed, is completely determined." Turing thus acknowledged to find himself engaging in the "age-old controversy" of free will and determinism. He suspected that "the feeling of free will which we all have" may be "an illusion," as if based on some partially random mechanism in the mindbrain. But we do not know and may not be able to know. So, in any case, given our views and biases as humans, Turing concluded that for the machine to pass as actually imitating a brain it "must appear to behave as if it had free will." He even discussed the possibility of introducing a random element into the machine, but dismissed it as a mandatory requirement:

> It is, however, not really even necessary to do this. It is not difficult to design machines whose behaviour appears quite random to anyone who does not know the details of their construction. Naturally enough the inclusion of this random element, whichever technique is used, *does not solve our main problem*, how to programme a machine to imitate a brain, or as we might say more briefly, if less accurately, to think. But it gives us some indication of what the process will be like. We must not always expect to know what the computer is going to do. (TURING, 2004 [1951], p. 485, emphasis added)

I take from this passage that Turing reduced (i) the problem of how to make a machine to appear to behave as if it had free will to (ii) the problem of how to program a machine to imitate the brain, or to think. Intelligence must enable the machine to behave in such a way that we cannot always expect to know what it is going to do. Turing sorted out the problem of making an observer to construe free will in the behavior of the machine, I propose therefore, by returning to his view of intelligence as an emotional concept. Turing thus reduced free will to intelligence — which has to appear partially random anyway —, for he was not even sure that free will has a real basis in the brain.

Indeed, whatever Turing's approach to problem (ii) is, let us take it to correspond to Turing's ontological concept of thinking. Now, it is not hard to see that the two concepts, the ontological (unobservable, but objective and mathematical) and the epistemological (observable, but subjective and emotional), must then be strongly coupled to one another by design. Together they constitute, I hold, Turing's scientific view of thinking or intelligence. The former is a hypothetical construct, but it is tied up to the latter to form an empirical or evidential basis for machine intelligence. This basis is, of course, implied in the Turing test (§3). But it is not meant

to be employed crudely. It is subject to the (idealized) condition that trickery or cheating is not allowed, as we are about to see (§2.4).

Towards Turing's ontological concept of thinking

Now, if Turing's approach for how to tackle problem (ii) is informed by the interaction between imagining intelligence and construing free will, what could that be? Turing said to believe that "the process should bear a close relation of that of teaching" (2004 [1951], p. 485). As known, Turing discussed that at length in section §7 of his (1950) paper. And he actually resumed his core May 1951 passage quoted above by saying:

> We should be pleased when the machine surprises us, in rather the same way as one is pleased when a pupil does something which he had not been explicitly taught to do. (TURING, 2004 [1951], p. 485)

Indeed, Turing found in machine learning a general approach to the conceptual problem of how to make a machine to think. This is because, for Turing, a machine that can learn both from teaching and from experience in general will, of course, be able to surprise us and can learn to cause wonder as but one of many competencies to be learned. As we shall see next, for Turing, I interpret, learning from experience was the way to address the requirement that wonder can never cease.

2.4 Learning: Turing's ontology of thinking

Right in the beginning of the 1952 BBC roundtable, Richard Braithwaite opened by acknowledging the paradoxical aspect of the question "can machines think?," and then asked Jefferson to start the discussion by commenting on the concept of thinking. Jefferson gave a relatively long reply and eventually converged onto his notion of thinking as the sum total what the brain of man or animal does, when he then asked Turing to take over:

> *Jefferson*: [...] One might say in the end that thinking was the general result of having a sufficiently complex nervous system. Very simple ones do not provide the creature with any problems that are not answered by simple reflex mechanisms. Thinking then becomes all the things that go on in one's brain, things that often end in an action but don't necessarily do so. I should say that it was the sum total of what the brain of man or animal does. Turing, what do you think about it? Have you a mechanical definition?
> *Turing*: 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. The important thing is to try to draw a line between the properties of a brain, or of a man, that we want to discuss, and those that we don't. To take an extreme case, we are not interested in the fact that the brain has the consistency of cold porridge. We don't want to say 'This machine's quite hard, so it isn't a brain, and so it can't think.'
> (TURING et al., 2004 [1952], p. 494-5)

For Turing, therefore, thinking is some vague mental process that goes on in the head of the thinking entity. Note that although this may not say much, it actually does suggest, as opposed to in his 1948 report, that this time (in 1952) he did not focus on thinking as an attributed property by an external observer but rather on some really existing, observer-independent process. In fact, in passages that we have already seen (§2.3), Turing have given away related hints about his view. He had said that the extent to which we regard something as behaving in an intelligent manner is determined in part "by the properties of the object under consideration," and he suggested that there is "the cause and effect working themselves out in the brain" which, if seen by the observer, tends to be regarded as thinking no more. So, we shall now ask, what did Turing consider to be the source of that thinking process that goes in the brain?

It turns out that Turing did have an ontological concept of thinking, and it consists essentially of learning like a brain, from experience. And this, in Turing's view, I hold, did not boil down to behavioral disposition only. It was meant to be grounded (have referent) in the machine as *alterations in its instructions or program*, just as it does in the (human) brain, Turing considered, as alterations in its neural structure. This is, I claim, precisely what Turing meant by "to imitate a brain, or as we may say more briefly, if less accurately, to think" (2004 [1951].p. 485). Note that it does not even takes a stand in the historical divide between symbolic v. connectionist approaches to artificial intelligence. As known, Turing did not dismiss the "symbolic language" approach — as long as it involves learning in the above sense and tolerates uncertainty, for "the processes of inference used by the machine," Turing wrote, "need not be such as would satisfy the most exacting logicians" (1950, p. 458). In fact, a (probabilistic) logic program can be learned from experience as well, although the analogy with the structure of the brain in the connectionist approach is stronger. To provide support for my claim, I will dedicate the rest of this section to present chronologically an extensive primary-source textual basis on Turing's most assertive statements about what is needed to make a thinking machine.

At least as early as his December 1945 report to the National Physical Laboratory (NPL, §A.3.2), Turing referred in print to machine "intelligence," which he considered to be possible if the machine is allowed to make mistakes:

> [...] 'Can the machine play chess?' It could be fairly easily be made to play a rather bad game. It would be bad because chess requires intelligence. We stated at the beginning of this section that the machine should be treated as entirely without intelligence. There are indications however that it is possible to make the machine to display intelligence at the risk of its making occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess. (TURING, 2005 [1945], p. 389)

This connection between intelligence and error was far from obvious back then, and would be indeed crucial in the development of Turing's views on machine intelligence (§A). It will reappear as we proceed in this chronological review just in connection with *learning*. Turing had

observed ever since his wartime service in Bletchley Park, it turns out, that learning is a heuristic process and as such it is hardly possible if errors are not allowed.

There is also an interesting related testimony by Donald Bayley, who worked with Turing during the war between 1943 and 1946. When he joined the NPL, Bayley related, Turing said that he would build "a brain" (Cf. SYKES, 1992, 25-27'; Cf. COPELAND, 2004, p. 374). This is also confirmed by both an October 1946 event that gave public exposure for Turing's view of the ACE as (Britain's) "electronic brain" (§A.3.3), and a *c.* November 1946 letter from Turing to cybernetician Ross Ashby, where the idea of making the machine to imitate the brain would become more apparent. In this (1946) letter to Ashby, Turing informed that "[i]n working on the ACE" (the universal computing machine of which he was leading the design), he was "more interested in the possibility of producing models of the action of the brain." Turing proceeded to distinguish two kinds of brain imitation, one entirely "disciplined" and devoid of originality and the other with more "initiative" (this term would actually appear explicitly only in 1948). This is how Turing described the first kind, which would be similar to "the action of the lower centres" of the brain:

> The ACE [as a universal Turing machine] will be used, as you suggest, in the first instance in an entirely disciplined manner, similar to the action of the lower centres [of the brain], although the reflexes will be extremely complicated. The disciplined action carries with it the disagreeable feature, which you mentioned, that it will be entirely uncritical when anything goes wrong. It will also be necessarily devoid of anything that could be called originality. (TURING, 1946)

While describing the second kind, it is worth noting that Turing referred specifically to trying out variations of behavior and to changing neuron circuits by the growth of axons and dendrites:

> There is, however, no reason why the machine should always be used in such manner: there is nothing in its construction which obliges us to do so. It would be quite possible for the machine to try out variations of behaviour and accept or reject them [...] and I have been hoping to make the machine do this. [...]
>
> Thus, although the brain may in fact operate by changing its neuron circuits by the growth of axons and dendrites, we could nevertheless make a model, within the ACE, in which this possibility was allowed for, but in which the actual [hardware] construction of the ACE did not alter, but only the remembered data, describing the mode of behaviour applicable at any time. [...] (TURING, 1946)

At that point he had not yet used the word "learning" explicitly. He started to do it in his January 1947 lecture to the London Mathematical Society. The connection he established with "intelligence" is striking:

> One can imagine that after the machine had been operating for some time, the instructions would have altered out of all recognition [...] Possibly it might still be getting results of the type desired when the machine was first set up [...]. In such a case one would have to admit that the progress of the machine had not been foreseen when its original instructions were put in. It would be

> like a pupil who had *learnt* much from his master, but had added much more by his own work. When this happens I feel that one is obliged to regard the machine as showing *intelligence*. As soon as one can provide a reasonably large memory capacity it should be possible to begin to experiment on these lines. [...] What we want is a machine that can *learn* from experience. The possibility of letting the machine alter its own instructions provides the mechanism for this. (TURING, 2004 [1947], p. 393, emphasis added)

A few months after that, Turing decided to go for a sabbatical leave from the NPL. His intentions were registered by the NPL director in an official letter:

> As you know Dr. A. Turing [...] is the mathematician who has designed the theoretical part of our big computing engine. [...]
>
> He wants to extend his work on the machine still further towards the biological side. I can best describe it by saying that hitherto the machine has been planned for work equivalent to that of the lower parts of the brain, and he wants to see how much a machine can do for the higher ones; for example, could a machine be made that could learn by experience? [...] (DARWIN, 1947)

Here appears explicitly, in connection with Turing's letter to Ashby, his concern with learning by experience but now in imitation of the "higher" parts of the brain.

Turing further articulated this ontological concept of thinking based on learning in his 1948 *Intelligent machinery* report tot he NPL. Thereby, he described "the direct method" to implement machine intelligence(p. 429-30), which is to make a machine to reproduce under full discipline other (special-purpose) machines that are given to it. The first machine is thus used as a (general-purpose) universal Turing machine. The "other method," as he called, is to use a combination of "discipline" and "initiative" for educating a machine. In this case, the machine would start fully unorganized — and to implement this Turing sketched a connectionist mathematical model — and learn to perform each special-purpose function in a flexible way, trying "to bring both discipline and initiative into it at once" (p. 430). Turing's conceptual observations along these lines can be found in sections §§7, 12 ("Education of machinery," and "Discipline and initiative") of his 1948 report. To emphasize the central point, I will just quote from Turing's own "[s]ummary" at the end of the report:

> The possible ways in which machinery might be made to show intelligent behaviour are discussed. The analogy with the human brain is used as a guiding principle. It is pointed out that the potentialities of the human intelligence can only be realised if suitable education is provided. The investigation mainly centres round an analogous teaching process applied to machines.
> (TURING, 2004 [1948], p. 431-2)

Turing thus established his research agenda for the process of machine teaching (learning) as the objetive route towards making a machine to show human-level intelligence.

Turing's view of machine thinking as the capability to learn for itself from experience by altering its own instructions seems to have reached maturity in 1950. In fact, the very structure

of his 1950 paper gives emphasis to his ontological concept of thinking based on learning, for his research agenda to build intelligent machines meant specifically to build "learning machines." (We will further examine this point in §3.3.) Again he suggested that the best way to address Lady Lovelace's objection was by machine learning. He cited Douglas Hartree (1949) who quoted from Lady Lovelace but added: "This does not imply that it may not be possible to construct electronic equipment which will 'think for itself', or in which, in biological terms, one could set up a conditioned reflex, which would serve as a basis for 'learning"' (p. 70). Turing then wrote: "[t]his whole question will be considered again under the heading of learning machines" (1950, p. 450). And Turing clearly stated that his goal was to describe how to program learning machines to play well the imitation game, which as known he proposed to substitute the question whether machines can think. He thereby (in his section §7 on learning machines) even connected the notion of "mind:"

> Our problem then is to find out how to programme these machines to play the game. [...] Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. [...] (TURING, 1950, p. 455)

Turing thus outlined some ideas on how to program the machine to learn for itself. Since then, he referred to machine learning as the way to go in order to program a machine to imitate the brain in all of his remaining communications on machine intelligence. He developed it notably in his *c.* 1951 BBC radio lecture. He thereby crucially addressed the connection between Turing's epistemological and ontological concepts of thinking. Turing observed that learning was the general approach for the objective part, also because once a machine is really able to learn from experience, the capability to cause wonder (also required for intelligence) would follow as a matter of course:

> It is clearly possible to produce a machine which would give a very good account of itself for any range of tests, if the machine were made sufficiently elaborate. However, this again would hardly be considered an adequate proof. Such a machine would give itself away by making the same sort of mistake over and over again, and being quite unable to correct itself, or to be corrected by argument from outside. If the machine were able in some way to 'learn by experience' it would be much more impressive.
> (TURING, 2004 [*c.* 1951], p. 473)

Also worth noting is that Turing anticipated very specifically the risk of trying to meet his notion of machine learning by just satisfying some (narrow) metric. He posed most clearly that his approach towards making the machine to pass an intelligence test involves its education from experience in general, and passing the test should be just a consequence of such process, and not a goal in itself. Otherwise, Turing posited, it would be a gross form of cheating:

> This process could probably be hastened by a suitable selection of the experiences to which it was subjected. This might be called 'education'. But here we

> have to be careful. It would be quite easy to arrange the experiences in such a way that they automatically caused the structure of the machine to build up into a previously intended form, and this would obviously be a gross form of cheating, almost on a par with having a man inside the machine.
> (TURING, 2004 [*c.* 1951], p. 473)

Turing resumed it and affirmed even further what he meant (2004 [*c.* 1951]): "[a]s I see it, this education process would in practice be an *essential* to the production of a reasonably intelligent machine within a reasonably short space of time" (p. 473, emphasis added). He completed: "[t]he human analogy alone suggests this."

Again in May 1951 Turing revisited the problem of how to program a machine to imitate a brain or to think, as we have already seen to some extent (§2.3). In connection to learning from experience as the method that Turing had chosen, it is important to emphasize that he made a disclaimer about his ideas on education. He suggested that such ideas were outlined as a best effort only, given his limited knowledge of how learning takes place in the child's mind. He posed that any process could potentially do it, as long as it involved machine teaching (learning):

> I will not attempt to say much about how this process of 'programming a machine to think' is to be done. The fact is that we know very little about it, and very little research has yet been done. There are plentiful ideas, but we do not yet know which of them are of importance. [...] I will only say this, that I believe the process should bear a close relation of that of teaching.
> (TURING, 2004 [1951], p. 485)

Turing also discussed machine learning at length in the January 1952 BBC roundtable, and referred to it yet briefly in his *c.* late 1952 text on chess. The one remark of his that will add to what we have seen so far, namely, his expectations and hopes about the possibility of making a machine to *learn how to learn*, or as he suggested, "snowball" learning. For Turing, that would be an important sign about whether or not the research on machine learning is going in the right direction. He again referred to such learning principles in reference to the human "mind:"

> *Turing*: [...] One will have to find out how to make machines that will learn more quickly if there is to be any real success. One hopes too that there will be a sort of snowball effect. The more things the machine has learnt the easier it ought to be for it to learn others. In learning to do any particular thing it will probably also be learning to learn more efficiently. I am inclined to believe that when one has taught it to do certain things one will find that some other things which one had planned to teach it are happening without any special teaching being required. This certainly happens with an intelligent human mind, and if it doesn't happen when one is teaching a machine there is something lacking in the machine. (TURING et al., 2004 [1952], p. 497)

Finally, also in the 1952 BBC roundtable, Max Newman made a comment (in the presence of Turing) in connection with Turing's test that may sound contradictory to his prohibition of cheating or trickery as we have seen above. Let me quote:

> *Newman*: [...] Of course, this is a dull, plodding method [brute-force table lookup], and you could improve on it by using a more complicated routine, but if I have understood Turing's test properly, you are not allowed to go behind the scenes and criticise the method, but must abide by the scoring on correct answers, found reasonably quickly. (TURING et al., 2004 [1952], p. 496)

Turing's comment on cheating, I think however, was sufficiently strong in that regard. And by the sum total of the textual basis that we have seen, I do think that he was clear in ruling it out any method that is *not* based on some form of learning. This was for him, I hold, a pre-condition for a machine to be claimed as having intelligence. For example, simple yet scalable data retrieval from a very large table may render machine behavior whose source is unknown to average observers and may deceive them to pass as thinking. But Turing seems to have considered that only learning machines would be eligible for a test.

Altogether I hope to have gathered enough evidence to establish that Turing had both an epistemological and an ontological concept of thinking or intelligence, respectively. The former — causing wonder on an observer — constituted Turing's evidential basis upon which machine intelligence is to be put to test. The latter — learning from experience like a brain — constituted Turing's actual model of a mechanical mind. Together, respectively, they give form to Turing's epistemology and ontology of the mechanical mindbrain.

Now, how exactly did Turing view the relationship between the actual (human) mindbrain and its logically possible mechanical model? What attitude did Turing hold towards it. We shall see it next.

2.5 Turing's realist attitude towards his hypothesis

As mentioned, Turing's May 1951 BBC radio lecture "Can digital computers think?" (§A.4) can be seen as the central source for interpreting his views relative to the problem of his philosophy of science of the mind. Turing opened it this way:

> Digital computers have often been described as mechanical brains. Most scientists probably regard this description as a mere newspaper stunt, but some do not. One mathematician has expressed the opposite point of view to me rather forcefully in the words 'It is commonly said that these machines are not brains, but you and I know that they are.' In this talk I shall try to explain the ideas behind the various possible points of view, though not altogether impartially. I shall give most attention to the view which I hold myself, that it is not altogether unreasonable to describe digital computers as brains.
> (TURING, 2004 [1951], p. 482)

Let us examine the two modes of speech used by Turing. On the one hand, he echoed himself in the voice of a fellow mathematician who would have said to him that *he knows* that digital computers are brains. On the other hand, he modulated his last sentence by the weaker form "it is not altogether unreasonable to describe digital computers as brains." How shall we read

into these two modes? I read that the second mode is phrased in a dialectic with the view of "[m]ost scientists," who considered the description under analysis — digital computers as brains — "a mere newspaper stunt." Since Turing disagreed, he could either adopt an equally aggressive attitude, or adopt a milder attitude. He opted for the latter, keeping his typically elegant public posture towards philosophical opponents. This is evidenced by an acknowledgement of his own bias ("shall try to explain the ideas" though "not altogether impartially"). The first mode in turn masked in wit what his own view was: digital computers can be described as mechanical brains.

Turing even stated shortly after in the same source: "In fact, I believe that [digital computers] could be used in such a manner that they could appropriately be described as brains" (2004 [1951], p. 482). And completed: "I should also say that '[i]f any machine can appropriately be described as a brain, then any digital computer can be so described'." Turing considered that this last statement would sound startling, and he wished to explain it. Essentially, he recalled, it can be shown to follow from the universality property of digital computers:

> A digital computer is a universal machine in the sense that it can be made to replace any machine of a certain very wide class. It will not replace a bulldozer or a steam-engine or a telescope, but it will replace [...] any machine into which one can feed data and which will later print out results. In order to arrange for our computer to imitate a given machine it is only necessary to programme the computer to calculate what the machine in question would do under given circumstances, and in particular what answers it would print out. [...]
>
> If now some particular machine can be described as a brain we have only to programme our digital computer to imitate it and it will also be a brain. If it is accepted that real brains, as found in animals, and in particular in men, are a sort of machine it will follow that our digital computer, suitably programmed, will behave like a brain. (TURING, 2004 [1951], p. 482-3)

Turing went further to acknowledge that this "involves several assumptions which can quite reasonably be challenged." In particular, "the machine to be imitated must be more like a calculator than a bulldozer." That is, the focus is on intellectual activity of a brain, which Turing cast as "mechanical analogues of brains" (p. 483). Besides this assumption of scope (intellectual only), Turing also considered the need for the machine (brain) to be imitated to have a predictable behavior. But of this, as we have seen (§2.3), Turing found his way out in connection with the notion of free will. It was the appearance of having free will that mattered, and Turing reduced it into an aspect of the brain imitation program (as we have seen, the capability to learn from experience). Turing also considered two hardware technology requirements: sufficient speed and storage. He dismissed speed, which he pondered was already good enough even for imitating a human brain. But storage was a serious bottleneck back then, he thought:

> Our present computers probably have not got the necessary storage capacity, though they may well have the speed. This means in effect that if we wish to imitate anything so complicated as the human brain we need a very much larger machine than any of the computers at present available.
> (TURING, 2004 [1951], p. 483)

Turing then gave his answer to what could be in principle a very important computer architecture question from the point of view of his hypothesis on the imitation of the human brain. He judged that, in terms of hardware, only storage capacity was needed for his hypothesis to be to feasible:

> It should be noticed that there is no need for there to be any increase in the complexity of the computers used. If we try to imitate ever more complicated machines or brains we must use larger and larger computers to do it. We do not need to use successively more complicated ones. This may appear paradoxical, but the explanation is not difficult. The imitation of a machine by a computer requires not only that we should have made the computer, but that we should have programmed it appropriately. The more complicated the machine to be imitated the more complicated must the programme be.
> (TURING, 2004 [1951], p. 483)

So, any argument for the growth of machine intelligence, for instance, that a machine will pass the Turing test in some predicted date in the future, on grounds of computer power continuing an exponential growth trend, has nothing to do and cannot be associated with Turing's views.

We thus arrive at what I take to be the central passage in Turing sources that reveal his philosophical attitude towards the hypothesis on the imitation of the human mindbrain:

> In view of this it seems that the wisest ground on which to criticise the description of digital computers as 'mechanical brains' or 'electronic brains' is that, although they might be programmed to behave like brains, we do not at present know how this should be done. With this outlook I am in full agreement. It leaves open the question as to whether we will or will not eventually succeed in *finding* such a programme. I, personally, am inclined to believe that such a programme will be *found*. (TURING, 2004 [1951], p. 484, emphasis added)

Now, given also all the other related passages we have seen so far, I think that Turing's talk of "finding" can be best understood at face value. I interpret that Turing kept an attitude of scientific realism towards his hypothesis on the *existence* of a program for the *imitation* of the human brain. Let us examine clearly what existence and imitation shall mean here.

On the one hand, about the "existence," recall that earlier on, in his October 1949 exchange with Polanyi, Turing had already suggested that in his view the mind is specifiable in terms of some logic or program operated by the real brain. That is, he believed, I interpret, that a mechanical mindbrain exists within the human head. (Whether this real mechanical mind could account for the whole mind is another question which I will address next shortly.) For now, let us just assume that there exists in the human head, so Turing considered, some mechanical mindbrain as part of the real mindbrain. On the other hand, about the "imitation," recall also that Turing conceded to Polanyi that the discovery of the real mechanical mind (*i.e.*, the real program operated by the brain) was an impossible empirical task to solve. It would entail reverse engineering it by observation of its behavior alone, which would be an intractable problem. And in any case, Turing considered the nervous system in humans and animals to be "a continuous machine" (1950, p. 451). As such it could not be literally a digital computer, which is rather

"a discrete-state machine." So the real mechanical mindbrain could not be actually reproduced but only mimicked. So, what program did Turing believe that would be "found"? From all we have seen, I posit that it was *a model or digital replica of the real mechanical mindbrain*, which in Turing's view could approximate the real mechanical mindbrain very closely such that their performances in a wide class of most impressive intellectual tasks would be indistinguishable. Note that it is precisely because the ontological status of the mechanical mindbrain digital replica (or machine thinking) from Turing's point of view is different than actual identity with the real mechanical mindbrain (or human thinking) that the correct reference to it, I hold, is by a notion of $thinking_2$, rather than by $thinking_1$, as I suggested early on in the Introduction. Indeed, Turing did *not* set out to find the exact mechanical mind that he thought to exist under operation of the human brain. He targeted "imitating" it in a representative class of intellectual tasks. (How wide this class would be relates to the relative scope of the real mechanical mind as part of the real mind, which as mentioned we shall see next shortly.) It shall be clear at this point that Turing's appeal to imitation (or behavioral dispositions) is due to the empirical impossibility of having any direct access to what he had as the real mindbrain, and has nothing to do with behaviorism. Turing believed that such model of the mechanical mindbrain, once found, could be inserted into a real digital computer as a program. The program could then be fed by machine teaching with behavioral patterns from our culture. The computer would eventually, therefore, be able to imitate a human brain, "or as we might say more briefly, if less accurately," Turing posited, "to think" (2004 [1951], p. 485). Turing's belief may not come true but shall not be seen as superstitious if one considers that its basis was actually the possibility of discovering a mathematical model of human learning by the study of both the actual physiology of the brain and the actual educational development of the human child. Achieving the target imitation would be empirical evidence that could bring back new knowledge about the the real mindbrain. Passing his test, Turing thought, as I interpret, would in principle be an interesting existence proof of his hypothesis. (Nonetheless, to anticipate §3 a bit, the test will only be able to be sensibly applied when an adequate model of the mechanical mind had already been found. And by that time, the test would no longer be necessary as such.) Indeed, Turing related his test with his hypothesis in the passage just next to the above:

> I think it is probable for instance that at the end of the century it will be possible to programme a machine to answer questions in such a way that it will be extremely difficult to guess whether the answers are being given by a man or by the machine. I am imagining something like a viva-voce examination, but with the questions and answers all typewritten in order that we need not consider such irrelevant matters as the faithfulness with which the human voice can be imitated. This only represents my opinion; there is plenty of room for others. (TURING, 2004 [1951], p. 484, emphasis added)

Turing's connection of the test with his hypothesis is critical to identify his philosophical attitude. For the test that Turing thought out is not any test, but one that allows for continued intervention on the model of the hypothetical mechanical mindbrain. If it is really a replica of the real

mechanical mind, which means that it can approximate the typical human behavior really well, then it must withstand relentless question-answering intervention by the inquirer as long as desired. Turing chose the "question and answer method," which as we have seen (§3.3) was the socratic dialectical method to make the truth to come up from an interlocutor in conversation analogously to a midwife that gives birth to a child.

Now, let us address the question of the extension and limits of the real mechanical mindbrain as part of the real mindbrain from the point of view of Turing. It turns out that earlier on, in 1950, Turing clarified how he saw it by means of his "skin of an onion" analogy:

> The 'skin of an onion' analogy is also helpful. In considering the functions of the mind or the brain we find certain operations which we can explain in purely mechanical terms. This we say does not correspond to the real mind: it is a sort of skin which we must strip off if we are to find the real mind. But then in what remains we find a further skin to be stripped off, and so on. Proceeding in this way do we ever come to the 'real' mind, or do we eventually come to the skin which has nothing in it? In the latter case the whole mind is mechanical. (It would not be a discrete-state machine however. We have discussed this.) (TURING, 1950, p. 454-5)

I have consolidated in Figure 2 a schematic view, as I interpret, of Turing's "skin of an onion" image of the mechanical mindbrain. Its status, Turing posited, could turn out to be revealed either to be a proper part of the real mind or to be equivalent to it. In the latter case, Turing announced, the whole real mind is mechanical. Turing's uncertainty about the existence of a remaining non-mechanical core in the real human mindbrain is critical in relation to Newman's distinction as we have seen early on in the Introduction. In case there is such a non-mechanical core in the human mindbrain, there would be certain functions or operations of the mindbrain, perhaps those Newman referred to as "poetical, reflective," that would be beyond the reach of machines.

Turing's views on consciousness

Now let us go back to the October 1949 seminar in Manchester (§A.4.2), when at some point Turing is reported to have said:

> TURING: declares he will try to get back to the point: he was thinking of the kind of machine which takes problems as objectives, and the rules by which it deals with the problems are different from the objective. Cf. Polanyi's distinction between mechanically following rules about which you know nothing, and rules about which you know.
>
> POLANYI: tries to identify rules of the logical system with the rules which determine our own behaviour, and these are quite different things.
> (TURING et al., 2005 [1949])

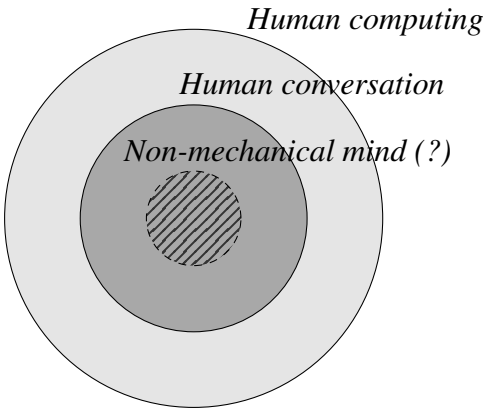

Figure 2 – Schematic view (cross-sectional) of Turing's "skin of an onion" image, as I interpret, of the mechanical mind as an outer part of the whole real mind. The first (outermost) skin would correspond to operations which we know that can be explained in purely mechanical terms (a human computer following strict rules and working with pencil and paper as Turing described in his 1936 paper). A second skin would correspond to operations that are typically said to require thinking or intelligence, most notably, keeping up a reasonable conversation as in Turing's test described in his 1950 paper. Other skin(s) would yet correspond to operations that may remain unexplainable mechanically, as if composing a hardcore of the " 'real' mind."

The reader may observe that Polanyi was posing the argument that John Searle reinvented and dressed in the Chinese room thought experiment. And Dorothy Emmet made a typical intervention in that connection, which lent us to have a spontaneous reply by Turing:

> EMMET: the vital difference seems to be that a machine is not conscious.
>
> TURING: a machine may act according to two different sets of rules, e.g. if I do an addition sum on the blackboard in two different ways:
>
> 1. by a conscious working towards the solution
> 2. by a routine, habitual method
>
> then the operation involves in the first place the particular method by which I perform the addition – this is conscious: and in the second place the neural mechanism is in operation unconsciously all the while. These are two different things, and should be kept separate. (TURING et al., 2005 [1949])

For Turing, the act of choosing a method to accomplish a task can be seen as an example of the exercise of consciousness. Also implied in that reply, I interpret, Turing thought that machines *can* afford self-referential (or "self-awareness") operations. The seminar conversation went on until arriving at Turing and Polanyi's exchange as we have seen in the beginning of this chapter:

> TURING: feels that this means that *my* mind as *I know it* cannot be compared to a machine.
>
> POLANYI: says that acceptance as a person implies the acceptance of unspecified functions. (TURING et al., 2005 [1949], no emphasis added)

In fact, Turing held a naturalized view of mind and consciousness, aside from the early modern tradition of founding knowledge on subjectivism. Polanyi seems to have tried to make a social and ethical point over Turing's epistemological and ontological discussion. Turing would return to a discussion of consciousness and self-awareness in machines in his 1950 paper, when discussing "[t]he argument from consciousness" and the "[a]rguments from various disabilities."

As known (cf. §A.4.1), Jefferson had outlined in June (1949a) some demands in order to accept that "machine equals brain." Turing, in his most explicit 1950 reply to it, cast Jefferson's demands as solipsism. He wrote: "[a]ccording to the most extreme form of this view the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking" (p. 446). Thereby Turing seems to have rejected the relevance of raw feels or qualia for the sake of discussing machine thinking. He wrote: "according to this view the only way to know that a *man* thinks is to be that particular man" (p. 446). And yet, shortly after in another passage still in discussion with "[t]he argument from consciousness," he added:

> I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper. (TURING, 1950, p. 447)

How shall we read into such a negative statement? At a first glance it seems to be in conflict to Turing's other statements on machine consciousness, which suggest that he held it to be as true as in human beings. I shall take one more moment to go through Turing's final 1950 note on the topic before giving my answer.

In fact, Turing proceed into discussing the "[a]rguments from various disabilities," and came back to the exact same notion of machine self-awareness as he had discussed in the October 1949 seminar:

> The claim that a machine cannot be the subject of its own thought can of course only be answered if it can be shown that the machine has *some* thought with *some* subject matter. Nevertheless, 'the subject matter of a machine's operations' does seem to mean something, at least to the people who deal with it. If, for instance, the machine was trying to find a solution of the equation $x^2 - 40x - 11 = 0$ one would be tempted to describe this equation as part of the machine's subject matter at that moment. In this sort of sense a machine undoubtedly can be its own subject matter. It may be used to help in making up its own programmes, or to predict the effect of alterations in its own structure. By observing the results of its own behaviour it can modify its own programmes so as to achieve some purpose more effectively.
> (TURING, 1950, p. 449, no emphasis added)

Turing considered, as it seems, that the very notion of a machine that could learn for itself (it would have to look at and alter its own instructions) would imply already a form of self-awareness. So it must have seemed to him that consciousness as self-awareness was nothing but a corollary to thinking or intelligence. But there was a trap in it.

In the passage immediately following, Turing argued: "[t]he criticisms that we are considering here are often disguised forms of the argument from consciousness". He then proceeded to reiterate on his epistemological (subjective, emotional) concept of thinking, which he firstly articulated in 1948 as we have seen (§2.3). In effect, Turing pondered, whenever he tried to show examples of various machine capabilities (self-awareness included), interlocutors were unimpressed — "the method (whatever it may be, for it must be mechanical) is really" so they thought, "rather base". Turing's solution to this general issue was to propose an evidential basis for machine intelligence that would control for confirmation bias against machine capabilities by taking the observer by surprise. For this to happen, of course, the machine's response must lie within the reach of what the observer might possibly expect. Otherwise, it would not even be intelligible. Now, machine self-awareness, as any other intrinsically private capability, seems to violate just that. It is in this connection, I suggest, that we may interpret Turing's point that there is "something of a paradox connected with any attempt to localise [consciousness]" from the external observer's perspective. I think this is a possible explanation for the tension between Turing's acknowledgement that some "mystery" may remain left on consciousness, and his other comments as if this were all but straightforward. Not unrelatedly, Turing pushed Jefferson's demands (requiring feelings to qualify for thinking) back as solipsism.

Now, it is important to distinguish Turing's move here from materialism reductionist approaches to (human or machine) agency, actions, or even feelings and consciousness. His attitude differs significantly from that of reducing subjectivist talk on phenomena such as, say, feeling pain, into (C-fiber) neurophysiology phenomena. Were this not the case, why would Turing have come up with an empirical basis of intelligence as a subjective, emotional concept? This has been shown by Proudfoot in her rebuttal of the view of Turing as a behaviorist (2017b). What he did, in my view, is rather to rebut the conventional wisdom taken for granted that, ontologically, thinking or intelligence belongs to the subjectivist talk and cannot be shown to take place in mechanical terms. Turing thought that this does not make sense and neglects the capabilities of the new machines — specially, indeed, in case our own mindbrain basis is wholly mechanical.

It should have already become apparent at this point that Turing's target was less a philosophy and more a physics of the mind, indeed.

2.6 Nine possible interpretations on existential hypotheses

In a classical paper (1950), Herbert Feigl formulated nine possible interpretations of existential hypotheses in science. (I shall introduce what he meant by existential hypotheses soon.) He thereby dismantled the realism-phenomenalism (-positivism, -operationism) dichotomy into several shades of gray in search of a rapprochement. Feigl's goal back then, in the same year that Turing proposed his test, was to examine questions such as:

> (1) [...] Can the analysis of existential hypotheses and theoretical concepts in terms of postulate systems, coordinating and operational definitions be upheld in the light of the actual procedures of science and their underlying semantical presuppositions?
>
> (2) Can we avoid both the reductive fallacies of phenomenalism [or operationism etc.] and the redundancies and confusions of metaphysical realisms?
>
> (3) Does not the notion of the probability of existential hypotheses presuppose a 'realistic' frame which cannot itself be meaningfully justified by considerations of probability?
>
> (4) What is the proper logical form for the definition (introduction) of hypothetical constructs which cannot be introduced by explicit definition in terms of observables? [...] (FEIGL, 1950, p. 35)

Feigl dated these questions back to the foundational discussions of Bertrand Russell (2008 [1910], 2009 [1914]), and "later, more emphatically," of Rudolf Carnap (1966 [1928]). In particular, he singled out for his critical discussion Russell's "(by now classical) programmatic pronouncement of his 'supreme maxim of scientific philosophizing'; *Wherever possible, logical constructions are to be substituted for inferred entities*" (2008 [1910], p. 155). Feigl set out to discuss it in depth in his "characterization of the hypothetical super-structure of science" (p. 41). In what follows I shall rely upon his endeavor for a specific reason: I think it will provide a normalized language for the location of Turing's philosophy of science of the mind and its received views.

To introduce his notion of existential hypothesis, Feigl observed that in the empirical sciences there are scarcely any interesting cases of hypotheses that make sole and essential use of "unlimited existential quantifiers" (p. 42), say, as in "there exists a thinking machine". Typically in empirical studies, instead, the existential quantification is actually constrained in real-world examples, say, "this particular machine thinks." But descriptive singular cases, of course, can be revealing for and can entail the more general formula, say, "if this machine thinks" then "there is some machine that thinks."

Feigl distinguished between existential hypotheses (type A and type B) that assert (respectively) *directly* and only *indirectly testable* states of affairs. Here are his examples:

> A) There are some matches in this match box. There is oil underneath Houston, Texas. There is a brain in Lord Russell's head. There is organic life on Mars. There is a further planet beyond the orbit of Pluto.
>
> B) This rock contains uranium oxide. There is calcium vapor on the surface of the sun. This light beam is plane-polarized. This room is traversed by radio waves. John's lungs are infested with filter-passing viruses. John's brain contains memory-traces. [...] Electrons are concentrated on the surface of this copper bar. Protons are moving rapidly through this cloud chamber. (FEIGL, 1950, p. 43)

Turing's hypothesis — "this machine thinks," or "this machine contains a mechanical mindbrain" —, we may promptly class, I invite the reader to observe, as type B. It can only be tested indirectly.

Feigl's criterion of differentiation between direct and indirect testability referred to whether or not the outcome of the test can be decided by fairly ordinary sense perception. This,

Feigl argued, "may be acceptable at least as long as we are not too exactingly scientific in our demands for proper identification and as long as we don't bother ourselves epistemologically about the reliability of our perceptual or mnemic performances." In this sense, he observed for the type-A examples, "the presence of the asserted state of affairs is ascertainable or refutable as soon as an opportunity for unhampered observation of the specified spatio-(temporal) area is afforded." We shall assume for instance, Feigl suggested, "that we know oil when we see, touch or smell it." By "directly testable," he also meant that "the obstacles that may at the moment prevent the (direct) testing of the given hypothesis are of a merely practical or technical character." The relevant test is known to be possible, he argued, "precisely because it is the kind of test that has been performed on previous occasions (drilling a well, opening a skull) or that could be performed if the required occasion were afforded or the required means perfected (rocket ship for the trip to Mars, etc.)" (p. 43). Feigl went on to argue that one might try to question into his type A-B distinction on account that the difference between the two groups of examples is merely one of degree. In some sense, it might be urged, the notion of ordinary sense perception is just what may be considered to be extended by the use of scientific instruments such as telescopes, optical (or electron-) microscopes etc., which would fill the gap, if there was any at all, between the directly testable and the only indirectly testable hypotheses. But this, Feigl argued, is just what foreshadows in a very rough and preliminary form the main issue of his discussion.

Confirmation rules for existential hypotheses

Quite clearly, Feigl posited, "any confirmation [...] of existential hypotheses (type A or B) must make use of a *confirmation rule*" (p. 44, my emphasis). And confirmation rules are proposed as laws (functional relationships) that relate properties (magnitudes).

For existential hypotheses of type A, only directly observable properties are related to one another. For example, the anatomical structure of crows as related to the black pigmentation of their feathers; or the presence of a backbone and the existence of a central nervous system; or more generally, anatomical laws of co-existence of structures and/or traits. We can think of such a law of co-existence as a confirmation rule, say, for the presence of a brain in a skull fossil. It is critical here, *n.b.*, that for type-A hypotheses the law may itself be confirmed, although never completely, by favorable instances. In justification of such reasoning, Feigl argued, "we should unhesitatingly quote the law (deterministic or statistical) that is drawn upon as a ground of validation." The more completely the law is confirmed, the higher the degree of confirmation for a given existential hypothesis on its basis.

Now crucially, Feigl asked, what about existential hypotheses of type B? How could they be confirmed by some law, if they comprise some only indirectly observable property, *e.g.*, an electron, or a magnetic field, or memory-traces, or a mind? Since these are *not* directly observable, on what basis can we establish the law itself, so that we can use it to confirm a particular instance of the indirectly observable properties? For example, Feigl considered, "we

are apt to say, glibly enough, that the deflection of a magnetic needle is evidence for the presence of a magnetic field." But how, he asked, "do we ascertain the validity of the law (type B) that relates the behavior of needles to magnetic fields?" Is this functional relationship a construction, an inference, a postulation? The existence and/or quality of the magnetic field is, *n.b.*, a type-B hypothesis, just like Turing's.

Not arbitrarily, Feigl chose this example in electromagnetism for the illustration of his discussion. He wanted to pick up a case "where the notion of *thinghood* is irrelevant" and then "confusion (engendered by misplaced picturizations) are more easily avoidable" (p. 36, no emphasis added). This observation can be just as interesting for the case of Turing's hypothesis. And the analogy must be straightforward: the deflection of the needle (directly observable) corresponds to interesting machine behavior in question-answering as demanded in Turing's test, and the presence of the magnetic field (only indirectly observable) corresponds to the presence of thinking or the mechanical mindbrain. We shall now be ready to go through Feigl's survey.

Feigl's nine possible interpretations of existential hypotheses

I reproduce below in some detail Feigl's surveyed answers, naive or sophisticate, abstracted from history to address his question about the validity of the law supposed to confirm the existence and/or quality of the magnetic field from the deflection fo the needle (1950, p. 44-50). In analogy with the example in electromagnetism, we may keep in mind Turing's hypothesis and the interpretations we have seen from the literature (§2.2):

1. (*Naive physical realism*). "Since the behavior of the needle must have a cause, we maintain that that cause (the magnetic field) exists even if this cause is not independently accessible to direct verification. The existence of the magnetic field is required by the principle of causality and is confirmed by the deflection of the needle" (p. 44).

2. (*Fictionalistic agnosticism*). "The observable behavior the needle displays, is in every respect as if there were an independently existing (but forever unknowable) reality: the magnetic field. The concept of the field is a useful fiction" (p. 44).

3. (*Probabilistic realism*). "The independent existence of the field cannot be asserted with certainty. But it can be inferred with probability from the behavior of the needle and other items of evidence. Quite generally, all inference that proceeds from observables to (directly) unobservables must be based on inductive probabilities [...]" (p. 45).

4. (*Naive conventionalistic phenomenalism, operationism, positivism*). "There is no conceivable way of independently and directly testing the existence of the field. Therefore to speak of a law (type B) relating the needle's behavior to the field, is extremely misleading. The concept of the field is *defined* by the behavior of the needle and has no meaning over and above what could be stated (more cumbersomely, to be sure) about the actual behavior of

the needle. 'Laws' of type B are nothing but definitions. Hypothetical constructs are thus regarded as strictly circular. Our hypotheses are so chosen that they parsimoniously (i.e. in more succinct language) summarize what could in principle be formulated as regards the *actually* observ*ed* facts" (p. 45, no emphasis added).

5. (*Critical phenomenalism, operationalism, positivism, or behaviorism*). "The tests for the presence of a magnetic field are not limited to the behavior of needles. A magnetic field 'manifests' itself also in the effects upon electric currents and upon the trajectories of electrically charged particles; in the rotation of the plane of polarization of light beams (Faraday, Kerr); in the Zeeman effect, etc. [...] The short-circuit circularity alleged by Naive Conventionalism [4] is thus avoided. [...T]he assertion of the existence of a particular magnetic field means (over and above the specific evidence that may have suggested the hypothesis) the total system of implicative relations between all sorts of conceivable test conditions and their corresponding test results. Hence the assertion (law of type B) that a needle is being deflected by a magnetic field (of given strength, direction, spatio-temporal extension) is far from tautological. It connects by synthetic statements all the various effects that would (within the specified spatio-temporal region) be observable on needles of all sorts and all the other effects (Faraday, Kerr, Zeeman, etc., etc.) that would occur in the same region under appropriately contrived conditions [...]" (p. 45-6).

6. (*Formalistic or syntactical positivism*). "A view which may well be regarded as a variant, or perhaps rather as an amplification, of the preceding one, focusses attention upon [...] postulates in a calculus which is so constructed that [...] all the empirical laws of a given field, e.g. electro-magnetics, are deducible from it; and which is interpreted via coordinating definitions. Either certain abstract, undefined concepts or else some concepts explicitly definable in terms of those primitives, are thus set in correspondence to the empirically or operationally defined constructs that have their place in the empirical laws. This requires a distinction either between empirical and theoretical constructs; or between empirical constructs and their mathematical idealization and formalization in a pure calculus. Although this view lends itself also to realistic interpretations, it is mentioned here as an important refinement of phenomenalism. As such it coincides fully with [5] and contributes additional plausibility to the view that the entities which figure in the laws of theoretical science are *nothing but* useful formal constructs; the theories themselves being "nothing but" mathematical models. The upshot then is still: the theoretical constructs are auxiliary devices, they are *façons de parler*, abbreviatory schemes for the description of the complex relationships between observables. This view with its emphasis upon the pure syntax of calculi, introduces a yet more distinctly nominalistic tinge into phenomenalism" (p. 46-7, no emphasis added).

7. (*Contextualistic phenomenalism*). "Another view, not too far removed from the two preceding ones, may be characterized as follows: Since the empirical constructs of science

(e.g. magnetic field strength, electric field strength, electric charge, intensity of electric currents, electromotoric force, conductivity, etc. etc.) are all linked together in a network of relationships, it depends upon the context of experimental investigations, and is in this sense somewhat arbitrary, *which* of these relationships may be regarded as genuine laws (synthetic propositions) and *which* others are then taken to be definitions (conventions, analytic propositions). Since in actual research laws of types B [...] and hypotheses of types B are never capable of test in isolation but always in the *context* of a whole system of relationships our initial query ('How are laws of type B to be validated?') is here considered as too simple-minded. In testing one hypothesis we invariably fall back on others which in this context are construed as definitions and provide the indispensable (and for the time being unquestioned) background and presupposition without which the very notion of a test of this kind is impossible." It can be cited as a classical example of this Newton's second law. It has defied attempts of formalization (say, in the spirit of David Hilbert's sixth problem, and of rational mechanics), as one cannot define whether force is primitive and mass theoretical, or the other way around. "[...] Now it is one of the essential features of the system that the relations as formulated in laws and hypotheses of type B *may* be taken as relations of causal or functional dependency, provided that other relations in the same system are accordingly interpreted as *definitions*. A certain surplus meaning for existential hypotheses of type B is thereby justified. Yet, if that surplus meaning is considered to be completely reducible to the directly testable (i.e. the evidential basis) the present view reveals itself as a variety of phenomenalism" (p. 47, no emphasis added).

8. (*Explanatory or hypothetico-deductive realism*). "An alternative interpretation of the methodological situation just described is found in a view that, I think, is very widely held, but only rarely explicitly stated: The 'dualistic' (i.e. realistic) assertion of the independent existence of the referents of hypothetical constructs is an essential and indispensable feature of any satisfactory explanatory system." Feigl quoted V. F. Lenzen's interpretation: "The dualistic theory of perception is based on the constructive hypothesis that perceptions are caused by independent things that radiate influences to the perceiving organism. Causality may be interpreted as a functional relationship between thing and percept, but even with this restriction the hypothesis is not capable of direct confirmation. It is confirmed by its success in explaining past perceptions and predicting future ones" (p. 47-8). I suggest the reader to observe the contrast with (1) naive physical realism, in which the existence of the magnetic field (the thing) would be inferred and thus promptly confirmed by the effect on the needle (percept). Here, in hypothetico-deductive realism, the field is independently conjectured to exist and its confirmation is achieved by the explanation of multiple percepts, past and future. Feigl himself proceeded to contrast it with interpretation (3): "[w]hile of course closely related to Probabilistic Realism [3], in its basic outlook, Explanatory Realism is not committed to the questionable justification by means of inductive probability. Considerations of probability are here, as in any case, indispensable when it comes to

the choice between different hypotheses. But the decision to supplement phenomenal description at all with 'transcendent' hypotheses is not in itself based upon inductive arguments. This view, however, provides only a hint, but no definite answer, as to the precise analysis of the asserted 'independence' or 'surplus meaning"' (p. 48). Feigl then arrived at the next interpretation, which he saw as a "semantical refinement" of this one.

9. (*Semantic or empirical realism*). "The missing explication [given by Wilfrid Sellars in a series of papers] has been advanced in semantical terms. The surplus meaning is understood to consist in the *factual reference* of the constructs employed in theoretical laws [of type B] and the existential hypotheses (of type B). This requires a clear distinction between *epistemic reduction* (i.e. the evidential basis) and the semantical relation of *designation* (i.e. reference). [...] We may say then that we must distinguish between the radical empiricist's meaning of "meaning" (i.e. epistemic reduction) and another, more common-sensical meaning of "meaning" (factual reference). [...] The very phraseology of indirect verification (confirmation) of statements requires for its explication a conceptual model in which statements as well as the states of affairs that render these statements true, can be represented. [...] 'Directly tested' likewise makes sense only if there is a theoretical model in which it is contrasted with and supplemented by 'indirectly tested.' [...] Only when we impose the requirements of pure pragmatics [as shown by W. Sellars] do we attain the desired scope of genuinely designating terms. That is to say, that in the language of empirical science all those terms (and only those terms) have factual reference which are linked to each other and to the evidential base by nomological relationships. Concepts or constructs that designate directly observable items of the world and those which do not, but are required for the coherent spatio-temporal-causal account to which science aspires [...] are thus properly related to each other by means of the metalanguage of pure pragmatics and semantics" (p. 48-50, no emphasis added). Now, a detailed review of Feigl's semantic realism is beyond my scope here. But one shortcut to it, I think, can be given — in contrast to (8) "explanatory realism." As a form of empiricism, Feigl's semantic realism eliminates the notion of a transcendental existence of hypothetical constructs such as the magnetic field. The latter's *existence* does not have to be *justified*, only *explained* within a coherent spatio-temporal-causal account of reality — where its existential meaning (factual reference) is linked to the meaning (factual reference) of directly observable constructs and to the evidential base by nomological (lawlike) relationships. In short, confirming the magnetic field does not entail its independent (transcendental) existence but its factual reference in a coherent causal account of reality. For additional interpretative boundaries and details on this view, I refer the reader to Feigl's (1950) text (cf., in particular, p. 50-2) and to a recent refresher by Matthias Neuber (2011).

Feigl posited that the order in which these nine interpretation classes were deployed was neither chronological nor systematic, but defined by "considerations of expository and dialectical

efficacy" (p. 52). Also, he added, "[f]rom a more systematic approach it is easily seen that:" (1) reappears in more sophisticated forms in (3) and again in (8) and (9). Likewise (2) and (4) may be absorbed in the more adequate formulations of (5), (6). Finally, much of (7) is compatible with and assimilable to (8) and (9).

We shall now be better positioned to revisit Turing's attitude towards his hypothesis and its received views.

2.7 Turing's received views revisited

Let us see how the interpretations on Turing's epistemology and ontology which we have seen before (§2.2) may fit into Feigl's interpretation classes. Except for a couple straightforward cases, I will not claim to class Turing's received views into Feigl's classes of interpretation exactly. But I do take the latter to provide us with an interesting language to normalize the discussion, and shall then revisit the former in their light tentatively.

The general view of Turing as a behaviorist associates him to Feigl's class (4) "naive conventionalistic phenomenalism." This view of Turing is very far away from the historical Turing and his proposal. This is to such an extent that it neglects even Turing's recurrent references to testing machine intelligence on the evidential basis of intellectual tasks (such as chess playing, code breaking, and mathematics) other than conversational question-answering (cf. 2004 [1948], p. 420-1; 1950, p. 460; and §A.3 in general), which give enough to upgrade his location into either of Feigl's classes (5) "critical phenomenalism" or (7) "contextualistic phenomenalism." I just rule out any chance to associate him with Feigl's class (6) "formalistic positivism" because even as a mathematician Turing rejected formalistic views such as Hilbert's program. For instance, he wrote to Max Newman in a (2004 [*c.* 1940]) letter: "I think you [Newman] take a much more radically Hilbertian attitude about mathematics than I do."

Moor (1976), as we have seen, found in Turing's proposal the goal of accumulating test results in terms of inductive evidence on the probability that machines can think or have thoughts as ordinarily understood. There should be little doubt that this view situates Turing's hypothesis and test within Feigl's class (3) "probabilistic realism."

Genova's (1994) interpretation in turn seems to have seen in Turing's proposal, as I quoted, the suggestion that "[g]ood simulation collapses the distinction between the real and the simulated" (p. 320). She understood computing from the point of view of Turing as "uninterested in the proper" and holding "no mirror to nature" at all. For her, Turing, as a homossexual, would have considered that natural differences all disappear "in the realm of numbers." She did find in Turing's proposal the view that "the only reason the machine can become human is because humans are already machines." But she seems to have interpreted not as a scientific hypothesis that was meant to be decided empirically, but as sort of a philosophical position towards Feigl's (2) class of "fictionalistic agonosticism."

Copeland's (2000b) interpretation, as we have seen, presented historical evidence against the view that Turing would have offered an operational definition of thinking or intelligence (p. 522-3). So it clearly rejects any association of Turing with Feigl's class (4) "naive conventionalistic phenomenalism," and likely with the other phenomenalism classes (5, 6, 7) as well. One might perhaps find a sign, as in Genova's view, of Feigl's (2) class "fictionalistic agnosticism," for in (2000a) he suggested that, for Turing, "it is 'not altogether unreasonable' to describe a machine that 'imitate[s] a brain' *as itself being a brain*" (p. 31, emphasis is mine). But Copeland sought to relate Turing's views with a physics of the mind. Accordingly, the only classes of Feigl's (if any) that remain are the realism ones (1, 3, 8 and 9).

Abramson (2011) and Proudfoot's (2013) interpretations, to the extent that they regard Turing's view of thinking or intelligence as *completely* based on "the epistemic-limitation condition" (Abramson) and on "an emotional concept" (Proudfoot) — and this extent is unclear to me —, they would imply to Turing an anti-realist philosophy of the mindbrain. It would be as if, for Turing, machine thinking or intelligence could only exist in case there is some external observer's mind to take notice. And this is a phenomenalist view, consistent with whatever of Feigl's classes (4, 5, 6, 7). Proudfoot's interpretation could also be seen as orthogonal to any ontological view and fully focused on Turing's proposal of an evidential basis for machine intelligence, but she seems to consider that any ontological theory of intelligence is inconsistence with viewing it also as an emotional concept (2017a, p. 295); whereas Abramson's, in turn, may be subject to another interpretation possibility. As we have seen, he cast the satisfaction of what he called Turing's "epistemic-limitation condition" as a necessary condition for the presence of a mind. He also argued that Turing's proposal was like Descartes's in that sense, although the latter was a dualist of mind-body and Turing was apparently a materialist. Now, if "mind" in Abramson's interpretation is nothing but an entity or property whose existence is just causally implied by satisfaction of the epistemic-limitation condition, then it would be suggestive of what Feigl called (1) "naive physical realism." Of course, advancing some specific interpretation of Turing's concept of "mind" in depth is no simple task, and this should be taken into account as a possible reason why commentators refrain from committing themselves to doing it.

All these views have historical and intellectual significance that go beyond my discussion here, and I profited from them. However, I have gathered extensive historical evidence (§§§2.3, 2.4, 2.5) to argue that Turing's philosophical attitude towards his epistemology and ontology of the mechanical mindbrain can be best associated with Feigl's (9) "empirical realism," as discussed next.

2.8 Turing's empirical realism on the mechanical mindbrain

Turing referred to his "main problem, how to programme a machine to imitate a brain, or as we may say more briefly, if less accurately, to think" (2004 [1951], p. 485). Let us now revisit, in

light of Feigl's scheme, what Turing meant by that. I have suggested before (§2.4) that, for him, the machine has to learn like a brain, from experience. And this, in Turing's view, does not boil down to behavioral disposition only. It was meant to be grounded (have referent) in the machine as *alterations in its instructions or program*, Turing considered, just as it does in the (human) brain as alterations in its neural structure. Now, note that the machine's table of instructions or program *exists*. It is directly observable, to such an extent that with a keyboard and a screen we can manipulate it. So, to think of alterations in a computer program shall not invoke any transcendental metaphysics but simply a physics. And yet, one may now ask: what does this all have to do with the real human mindbrain? In fact, to make it well grounded in history, Turing's most notable intellectual opponent Geoffrey Jefferson did ask just that in the very opening of his Lister-medal Oration "The mind of mechanical man:"

> No better example could be found of man's characteristic desire for knowledge beyond, and far beyond, the limits of the authentic scientific discoveries of his own day than his wish to understand in complete detail the relationship between brain and mind — the one so finite, the other so amorphous and elusive. It is a subject which at present awakes a renewed interest, because we are invaded by the physicists and mathematicians — an invasion by no means unwelcome, bringing as it does new suggestions for analogy and comparison. We feel perhaps that we are being pushed, gently not roughly pushed, to accept the great likeness between the actions of electronic machines and those of the nervous system. At the same time we may misunderstand this invitation, and go beyond it to too ready an affirmation that there is identity.
> (JEFFERSON, 1949a, p. 1105)

Later on in his lecture (1949a, p. 1110), Jefferson outlined his demands to accept that "machine equals brain." A reporter from *The Times* made a headline out of them, and submitted the problem to Turing by telephone on 10 June 1949. He replied sharply in wit, and this marked the origin of Turing's public stand on the question whether machines can think (§A.4.1). So there should be little doubt at this point that some notion of machine-brain identity was a conjecture within Turing's target. I suggest to call it intellectual-power identity, which comprises the intellectual properties of the (real human) mindbrain. This seems to be precisely what Turing meant (cf. his "skin of an onion" image in Figure 2) by the possibility of finding out that "the whole mind[brain] is mechanical." Now, let us see what this means according to Feigl's scheme.

Firstly, we know for sure of the existence of a mindbrain in the human head, from sort of a neutral monism point of view. Secondly, we also know by Turing's 1936 paper and the computer revolution related to it that some of its functions or operations can be explained in purely mechanical terms. Now, in connection with the possibility of a machine-brain identity, Turing's research question was how much of the real human mindbrain is mechanical and whether all of it (the whole mindbrain) is mechanical. In Feigl's terms, this will amount precisely to a type-B existential hypothesis: is the human mindbrain mechanical, or does the human head contain a (fully) mechanical mindbrain? Note that, while the real human mindbrain is directly observable, its casting as a *mechanical* mindbrain is only indirectly observable. So now

the problem will translate to asking: how could such a type-B existential hypothesis be ever confirmed?

The reader may recall that I had referred to *Turing's scientific hypothesis* before (§2.6) in two phrasing variants, namely, (h) "this machine thinks," or (h') "this machine contains a mechanical mindbrain." It turns out, however, that for Turing, if found true, the machine-based hypothesis would have to imply as a corollary its human-based version, which is nothing but the machine-brain identity. The latter is phrased ($h^\star$) "the real human mindbrain is mechanical," or ($h^{\star\prime}$) "the human head contains a mechanical mindbrain." Now, if Turing set out to pursue (h) and ($h^\star$), a key question to ask is: on what grounds? For Turing, it was critical whether or not a machine could accomplish the most impressive intellectual tasks one may think of. As we have seen (§3.5), Turing observed that our "higher" (intellectual) human capacities are often taken as empirical evidence of our higher place in nature — for example, as cast in Descartes' beast-machine thesis and Jefferson's Lister Oration. This was the essential argument that Turing mirrored in (h) and its corollary ($h^\star$). So, Turing thought, if indistinguishable evidence could be produced from machines then the same argument would have to apply to them, so runs the logic. Any impressive intellectual task would require human-level learning, whose real basis in (either) the (machine- or human-based) mechanical mindbrain can only be directly observed in the performance of the task itself. If achieved by a machine, this would be fairly reasonable empirical confirmation of his machine-based hypothesis (h), with the human-based counterpart ($h^\star$) following as corollary. Turing proposed his famous 1950 conversational question-answering test as tentative design for a most ambitious intellectual task, and does not seem to have been nitpicky about it (§3).

Overall, it should be now clearer, Turing proposed a type-B hypothesis on the existence of a real mechanical mindbrain in the human. In his view, it could be replicated (sufficiently approximated in a model) and built into a machine. The hypothesis confirmation was not supposed to be pursued by engaging any transcendent metaphysics at all. But that does not mean that he did not hold a realist attitude towards it. As I have shown through an extensive basis of primary-source passages, he referred (semantically) to "the functions of the mind or the brain," or "the real mind," or "[i]f now some particular machine can be described as a brain we have only to programme our digital computer to imitate it and it will also be a brain," or "I, personally, am inclined to believe that such a programme will be found." Other representative semantic references to the machine-brain identity made by Turing include his notion of a "state of mind" of a Turing machine (1936, p. 250); his thinking of the storage capacity of the machine as "memory" (2004 [1948], p. 413); his understanding of buying the machine "sense organs" (1950, p. 460). There are plenty.

Indeed, we have seen sketched (here, and in §§§2.3, 2.4, 2.5) the two essential components of Feigl's empirical realism, namely, the (semantic) factual referents of directly observable and only indirectly observable constructs that compose a coherent spatio-temporal-causal account

of reality, and their nomological relations as tied up to an evidential basis. A more detailed description of both goes beyond the scope of this dissertation and will be left for future work. I shall now though just add one element to it. Turing did a lot of thinking on how to bridge the gap between the two conceptual systems he conjectured identity for: brain anatomy and physiology, on the one side, and digital computers and Turing machines, on the other side. As we have seen (§2.5), he had no doubt that it was possible to make a digital computer to imitate a brain. So the core question he asked was not whether it is possible, but just *how* to make a digital computer to learn like a mindbrain? As of 1948, he considered two main approaches: discipline and initiative. In the approach centered on discipline, the learned competence of a universal Turing machine increases as the collection of special-purpose Turing machines it can imitates increases. In the approach centered on initiative, on the other hand, an unorganized machine having networked artificial "neurons" is taught everything from scratch like a child. Turing sought out combinations of the two approaches. As of 1950, he focused on and formulated the approach centered on initiative in terms of a child-machine program that would be taught everything from scratch by a schoolmaster (who is not its designer) and learn from experience (including going to school). In his 1950 paper, perhaps because the 1948 ideas were NPL's and classified Turing had not introduced the notion of an unorganized machine, but it seems that he was thinking of the child machine program as such. In this connection, Copeland and Proudfoot considered in (1996) that Turing was referring to his concept of unorganized machines when he wrote in (1950): "I have done some experiments with one such child-machine, and succeeded in teaching it a few things" (p. 457). Indeed, the historical evidence Copeland and Proudfoot gathered (1996) makes the case that Turing's notion of unorganized machines was the first proposal of a connectionist approach to artificial intelligence. Key ideas and concepts can be recovered from Turing's 1948 unorganized machines and his 1950 notion of educating child machines for the sake of a reconstruction of his argument and research agenda in pursuit of his hypothesis.

Progress by successive approximation versus *ignorabimus*

In his discussion on scientific realism, Feigl made a point which I see to have important connections with Turing's realism and the reaction of Jefferson and others to it. He wrote:

> The metaphysical realist craves for a "proof" of the existence of entities which are not directly verifiable. [...]
>
> The craving for direct verification seems cognate with the wish for immediate experience notoriously manifest on the deeper levels of epistemology in the camps of subjective idealism, radical empiricism and some of the older varieties of positivism. There the issue hinges upon two different notions of "reality"; (and, correspondingly, two different notions of "knowledge of reality"). One is the intuitive notion of reality — as stressed by Descartes, Berkeley and Bergson. The other is the empirical and scientific notion of reality. According to the first view the criterion of reality is direct experienceability. According to the second

> view reality is ascribed to whatever is required (confirmed) as having a place in the spatio-temporal-causal system. (FEIGL, 1950, p. 50-1)

It should be relatively straightforward at this point to find Jefferson (§2.5) attached to the first notion of (knowledge of) reality, and Turing committed to the second. Feigl had more to add:

> The danger of a related confusion may be seen in the perennially fashionable utterances of scientific agnosticism. "Even if we knew all *about* electricity (matter, life, mind) we should never know what electricity, (etc.) *really* is." Phrases of this sort (popular with great scientists, especially at the occasion of after-dinner-speeches, presidential addresses at association meetings) may be the expression of a proper and commendable humility in view of the tremendous and obviously incompletable tasks of scientific research. The phrase in this interpretation merely emphasizes that scientific progress is a matter of successive approximation. But frequently enough it is intended as a genuine "ignorabimus". No matter how complete our scientific knowledge, it would never acquaint us with the essence of things. This agnosticism could indeed be overcome only by such fanciful procedures as intuitive identification (perhaps real coalescence) of the knowing subject with the to-be-known-object. As long as our direct experience is limited to the data of our consciousness, we shall indeed never be able to "know" (by acquaintance) what electricity "really" is, because we should have to *be* an electric current in order to achieve that crowning feat of "real knowledge." It is truly astounding to find how widespread and deep rooted this confusion is as regards "knowledge" and how tenacious the wish for direct intuition. (FEIGL, 1950, p. 51-2, no emphasis added)

The closeness between Feigl and Turing's views — even at the level of their phrasings — is astonishing. It adds strength to the view that Turing's realism is akin to Feigl's. In (1967 [1958]), Feigl did not deny the reality of subjective experience or "knowledge by acquaintance" (as opposed to "knowledge by description," p. 79), which is expressed in the private or "solipsistic (egocentric)" language (p. 155). For an excellent account of Feigl's famous 1958 essay and 1967 postscript (*ibid.*), I refer the reader again to Thomas Neuber (2018 [2014]). Turing did not deny it either. He acknowledged it to such an extent that he took it as *the* evidential basis for machine intelligence (§2.3). However, Turing was not naive. He suggested the imitation game as a way out to neutralize biases of the cartesian "ego," in particular, the bias towards the conservative position that machines never thought, can't think and will never think, as posed by Jefferson.

Relative to Jefferson, in fact, Feigl's description of the "after-dinner" stripe of the "ignorabimus" motto could hardly have been more on target. In his (1949a) Lister Oration, Jefferson evoked humility in this fairly reasonable rhetoric:

> We must be ware of making science too rigid, self-conscious, and pontifical. A. N. Whitehead confessed to me once that he found that he had escaped from the certainty and dogma of the ecclesiastics only in the end to find that the scientists, from whom he had expected an elastic and liberal outlook, were the same people in a different setting. I am encouraged, therefore, to proceed in the hope that, although we shall not arrive at certainty, we may discover some illumination on the way. (JEFFERSON, 1949a, p. 1105)

I happen to have found out that Whitehead and Jefferson had just this conversation Jefferson alluded to in the occasion of an after-dinner conversation at Whitehead's house in New Haven on 20 May 1943. Jefferson's recorded notes of it are reproduced in full by his biographer Peter Schurr (1997, p. 231), and make up one page of joint criticism with the Whiteheads on certainty and stubbornness, having Bertrand Russell, Julian Huxley and David Hume as models.

Notwithstanding, let us recall that Turing was confronted by Jefferson with the view that our "higher" (intellectual) human capacities are often taken as empirical evidence of our higher place in nature. So, Turing thought, if indistinguishable evidence could be produced from machines then the same argument would have to apply to them. Accordingly, it would have either to upgrade the status of machines or downgrade the (higher) status of human beings. Turing seems to have focused on the former. The possibility of the latter, in any case, does not seem to have worried him, neither introspectively nor otherwise. In fact, he exposed his bold views on one of the biggest mass-media broadcasters worldwide. Somewhat in this connection, James McGilvray pondered:

> One possible explanation for Descartes's reluctance to venture into the mind by using the tools of science lies in Galileo's experience with the church. Descartes might have been unwilling to appear to be offering a naturalistic account of the mind, or what the church authorities might have thought the soul. (MCGILVRAY, 2009, p. 38)

Descartes's actual motives may lie in the past. But Turing's shall not, at least for some of us that seek the intelligibility of his views. As we have seen (§1), Turing had the scientific ambition of advancing our knowledge about ourselves (in the second sense articulated by Feigl) as quoted above. For just that, he seems to have had to cope with ecclesiastic-like "certainty and dogma."

Altogether, I hope to have contributed to show that Turing addressed the human mindbrain from a natural science point of view. But did he actually reduce the human personality to machine as well? It turns out that he did not, as discussed next.

2.9 Turing's views on the subjectivity of thinking machines

It is interesting to note that Polanyi scholar Paul Blum, having studied the Turing-Polanyi exchanges (2010), found that "Turing's approach, as documented here, does not lend itself to reductionism." Blum suggested this both in general and in particular. In general with respect to Turing's propositions in the discussion on mind and machine, meaning that he was careful enough to state his propositions within a defined scope. And in particular with respect to what became Polanyi's main object of study in his *Personal knowledge* (1974 [1958]), namely, the irreducibility of the personal attributes of a mind. For instance, in his (1939 [1938]) doctoral thesis, Turing wrote: "[m]athematical reasoning may be regarded rather schematically as the exercise of a combination of tool faculties, which we may call intuition and ingenuity" (p. 214).

And completed: "[t]he parts played these two faculties differ of course from occasion to occasion, and from mathematician to mathematician" (p. 215). This was posed in the context of his doctoral thesis, and what he suggested (§A.2.2) is that although Gödel's incompleteness theorems set limits to what can be achieved by a single machine, it established nothing about the powers of a series of machines, each one, Turing pondered, possibly as "personal" as an individual mathematician.

Turing did not reduce the personal mind to universal logical principles, nor did he reduce it to neurophysiology or whatever mechanism. As quoted in the beginning of this chapter, he posited to Polanyi that "the mind is only said to be unspecifiable because it has *not yet been* specified," but conceded: "this means that *my* mind as *I know it* cannot be compared to a machine" (2005 [1949], no emphasis added). So, any view that Turing would have been supportive of reducing people's raw feels or qualia to neurophysical principles of the brain is far fetched. Turing was no crass materialist. The evidence is suggestive, instead, that he would have been sympathetic to some normative (irreducible) concept of person. His dispute with Polanyi, though, and here comes again Turing's wits, seems to have been on not to deny the same to thinking machines. His plea was for applying a fair play to the machines.

In the eyes of Turing, I interpret, denying subjectivity to thinking machines was even contradictory, not for moral but for technical reasons. Let us see. Turing observed that there can be no learning from experience if not by allowing the learning entity to make mistakes. In fact, on "learning machines" in 1950 he wrote:

> Another important result of preparing our machine for its part in the imitation game by a process of teaching and learning is that 'human fallibility' is likely to be [imitated][1] in a rather natural way, i.e. without special 'coaching'. (The reader should reconcile this with the point of view on p. [449][2] .) Processes that are learnt do not produce a hundred per cent. certainty of result; if they did they could not be unlearnt. (TURING, 1950, p. 459)

(Note that this is yet another caveat made by Turing in connection with the interpretation of his imitation game. The machine to take a pretence in the test is not meant to be "coached" to pass it.) If learning from experience is the ontological basis of thinking or intelligence, then one shall not be able to avoid some consequences of having them. In fact, as an inevitable corollary of their learning, machines shall make their own mistakes and build (in a certain sense) their own subjectivity. If we shall never know "what is like to be a bat," neither shall we know what is like to be a machine. This is because, even if we actually have a mechanical mindbrain and a machine-based replica can be made to approximate it very closely, there is no reason to believe that neither machines in general (species-level) nor any machine in particular (individual-level) would happen to have the exact same raw feels as we do as species and as individuals.

[1] In Turing's text appearing in *Mind*, it was written here "omitted." Jack Copeland (2004) presumed it to be a typographical error (p. 463, note 6). The substitution for "imitated" is my own guess.

[2] This is where Turing discussed "(5) Arguments from various disabilities" and distinguished "errors of functioning" and "errors of conclusions."

Let us now recall from the Introduction Daniel Dennett's allusion to the social problem of the existence of thinking machines:

> It is of more than academic importance that we learn to think clearly about the actual cognitive powers of computers, for they are now being introduced into a variety of sensitive social roles, where their powers will be put to the ultimate test: in a wide variety of areas, we are on the verge of making ourselves dependent upon their cognitive powers. (DENNETT, 2006 [1984], p. 295)

The "thought" processes of learning machines are unexplainable by definition. It might be helpful to observe that Turing foresaw and offered important insights on such issues some seventy years ago. In Turing's obituary for the Royal Society, Newman related:

> The unexpected element in human behaviour he [Turing] proposed, half seriously, to imitate by a random element, or roulette-wheel, in the machine. This, he said, would enable proud owners to say 'My machine' (instead of 'My little boy') 'said such a funny thing this morning'. (NEWMAN, 1955, p. 255)

Are not we already seeing this to come true? In a rudimentary form, I think we do. What if Turing was right, and the true learning program that would really imitate the human mindbrain will be found?

Turing exposed himself on the mass media of his time to speak out on what he believed to be their true capabilities. He rejected the instrumentalist view of digital computers, which cast them in his time as instruments of industry and of war. Both from the scientific point of view of his hypothesis on the existence of the mechanical mindbrain and from the ethical and social point of view of his anti-chauvinism on gender and species forms, I think, there seems to be much yet to be learned from Turing.

2.10 Analytical summary

In this chapter I have presented Turing's scientific hypothesis on the mechanical mindbrain. I have strived for a clear-cut distinction between his epistemology and his ontology with hopes that this distinction would be fruitful towards a better understanding of both Turing's views in the primary literature and their reception in the secondary literature. I now offer a summary collecting my claims and key points hereby developed.

Received views (§2.2). Historically, Turing's philosophy of science of the mindbrain received a wide range of interpretations, from behaviorism to probabilistic realism, social epistemology, and so on. I reviewed representative views — besides the general readings of Turing as a behaviorist, I have reviewed in particular James Moor (1976), Judith Genova (1994), Jack Copeland (2000b), Darren Abramson (2011) and Diane Proudfoot's (2013) — and tried to read into their assumptions and conclusions on Turing's epistemology and ontology of the mindbrain. The views I have reviewed would be revisited later, as summarized below.

Turing's epistemology (§2.3). I have interpreted that wonder is the central concept of Turing's proposed empirical basis for machine thinking or intelligence so that we could test it intersubjectively. However learned a machine can be, there may be emotional barriers for one to attribute thinking or intelligence to it. Turing considered the attribution of intelligence to a given entity by an external observer does not depend only on "the properties of the object under consideration," but also ("as much") on "our own state of mind and training" as observers. He pointed out explicitly the importance of this subjective element in 1948, in 1950 and again in 1952. Turing seems to have observed the need to neutralize it in order to prevent or block confirmation bias towards the conservative position that machines never thought, can't think and will never think, as present in the position of Jefferson. It was based on this observation, I have suggested, that Turing set up a directly observable (external) empirical basis for intelligence. I also argued, nevertheless, that this should not be seen as exhaustive of Turing's concept of thinking or intelligence. If so, it would indeed be prone to the easy criticism that Turing would have proposed a research program to build a machine simply to deceive. Equipped with such a mathematically-empty notion, neither Turing nor anyone could get anywhere near building an intelligent machine. It would rather imply that Turing's research program would better have been to build a machine not to learn in general, but simply to deceive. But Turing himself rejected any resort to easy contrivances and, in fact, called it a gross form cheating. He also identified an interaction between the problem of making a machine to display intelligence or to think and the problem of making it appear as if it had free will. He sorted out it by reducing the latter to the former, for he was not even sure that free will has a real basis in the brain. Turing's scientific view of thinking or intelligence, I submitted, has two components. One is his epistemological concept of wonder (observable, but subjective and emotional), and the other is his ontological concept pf learning (unobservable, but objective and mathematical). They are strongly coupled to one another by design. Turing found in machine learning a general approach to the conceptual problem of how to make a machine to think. For Turing, a machine that can learn both from teaching and from experience in general will, of course, be able to surprise us and can learn to cause wonder as but one of many competencies to be learned. For Turing, I have interpreted, learning from experience is the way to address the requirement that wonder can never cease.

Turing's ontology (§2.4). I have interpreted that learning like a brain, from experience, is the central concept of Turing's ontology of thinking or intelligence, so that he could build it into a machine. The learning was to be grounded (have referent) in the machine as alterations in its instructions or program just as it does in the (human) brain as alterations in its structure through changing its neuron circuits by the growth of axons and dendrites. This, I have claimed, is precisely what Turing meant by "to imitate a brain, or as we may say more briefly, if less accurately, to think." And it does not takes a stand on the symbolic v. connectionist approaches to artificial intelligence. Turing did not dismiss the symbolic language approach, as long as it involves learning and allows for uncertainty. To provide support for this claim, I have presented an extensive chronological primary-source textual basis on Turing's most assertive statements

about what is needed to make a thinking machine. His statements along these lines date to at least as early as December 1945, and continued through *c.* November 1946, January 1947, summer of 1948, early 1950, 1951 and early until late 1952. In addition to the central notion that he was concerned with, namely, that learning from experience like a brain would be grounded in alterations in the machine's program without redesign, another key point on which Turing insisted was machine teaching. He thought that the analogy with the human brain was to be used as a guiding principle. But he made a disclaimer about his ideas on education. He suggested that such ideas were outlined as a best effort only, given his limited knowledge of how learning takes place in the child's mind. Turing also expected and hoped that it would be possible to make a machine to learn how to learn, what he called "snowball effect." For Turing, that would be an important sign about whether or not the research on machine learning is going in the right direction. Indeed, Turing considered that electronic computing machines should be planned not only for work equivalent to that of the lower parts of the brain, but also for the higher ones. He anticipated very specifically the risk of trying to meet his notion of machine learning by just satisfying some (narrow) metric. His approach towards making the machine to pass an intelligence test had to involve its education from experience in general, and passing the test should be just a consequence of such process, and not a goal in itself. He posited that it would otherwise be a gross form of cheating. Turing had both an epistemological and an ontological concept of thinking or intelligence. The former — causing wonder on an observer — constituted Turing's evidential basis upon which machine intelligence is to be put to test. The latter — learning from experience like a brain — constituted Turing's actual model of a mechanical mind. Together, respectively, they give form to Turing's epistemology and ontology of the mechanical mindbrain.

Turing's realist attitude (§2.5). Turing thought that digital computers can be described and used as mechanical brains. He related that with the universality property of digital computers. He then posited that if any machine can appropriately be described as a brain, then any digital computer can be so described. This would mean to program it to imitate the brain and then, Turing considered, it will also be a brain. If it is accepted that real brains as found in animals and in particular in men are a sort of machine, Turing pondered, it will follow that a digital computer, once suitably programmed, will behave like a brain. Turing acknowledged that these views involved several assumptions that could reasonably be challenged. In particular, the focus is on intellectual activity of a brain, which Turing cast as mechanical analogues of brains. He also considered the need for the machine (brain) to be imitated to have a predictable behavior. This was related with the notion of free will, and it was actually the appearance of having free will that mattered. Turing thus reduced it into an aspect of the brain imitation program, namely, the capability of displaying intelligence by learning from experience. Turing also considered two hardware technology requirements. Speed, he considered, was already good enough even for imitating a human brain. But storage he took to be a serious bottleneck back then. In terms of computer hardware and its architecture, Turing judged, only storage capacity was needed for

making a computer to imitate the human mindbrain. The hard scientific challenge that remained though was how to program the computer to do it. Turing conceded as an open question whether or not we will eventually succeed in finding such a program. He believed that it will be found. Also in light of all the other related Turing sources, I have proposed to understand his talk of "finding" at face value. I have interpreted that he kept overall a realist attitude towards his hypothesis on the existence of a program for the imitation of the human mindbrain. He considered that there is a real mindbrain in the human head, and a model could be specified and built into a digital computer to approximate very closely its mechanics. This model would not be the exact mechanical mind that he thought to exist under operation of the human brain. For one difference, Turing considered the nervous system in humans and animals to be a continuous machine, while digital computers to be in turn discrete-state machines. So the real mechanical mind could not be actually reproduced but only mimicked. I have pointed that Turing did not set out to find the exact mechanical mind that he thought to exist within the human brain. He targeted imitating it in a wide class of relevant intellectual tasks. I have pointed that Turing's appeal to imitation (or behavioral dispositions) is due to the empirical impossibility of having any direct access to what he had as the real mindbrain, and has nothing to do with behaviorism. The model or digital replica of the real mechanical mindbrain was to be close enough to it such that their performance in a wide class of intellectual tasks would be indistinguishable. Turing believed that such model of the mechanical mindbrain, once found, could be inserted into a real digital computer as a program. The program could then be fed by machine teaching with behavioral patterns from our culture. The computer would eventually, therefore, be able to imitate a human brain, or as we might say more briefly, if less accurately, Turing posited, to think. I have argued that Turing's belief may not come true but shall not be seen as superstitious if one considers that its basis was actually the possibility of discovering a mathematical model of human learning by the study of both the actual physiology of the brain and the actual educational development of the human child. Passing his test, Turing thought, would be an interesting existence proof of his hypothesis. (Nonetheless, I hope to have shown before, the test can only be sensibly applied when an adequate model of the mechanical mind had already been found. And by that time, the test would no longer be necessary as such.) Turing's connection of the test with his hypothesis is critical to identify his philosophical attitude. The test allows for continued intervention on the model of the hypothetical mechanical mindbrain. If it is really a replica of the real mechanical mind, which means that it can approximate the typical human behavior really well, then it must withstand relentless question-answering intervention by the inquirer as long as desired. Turing addressed the question of the extension and limits of the real mechanical mindbrain as part of the real mindbrain through his "skin of an onion" image. He believed that the actual status of mechanical mindbrain could turn out to be revealed either to be a proper part of the real mind or to be equivalent to it. In the latter case, Turing announced, the whole real mind is mechanical. Turing held a naturalized view of mind and consciousness, aside from the early modern tradition of founding knowledge on subjectivism. In this connection, Polanyi seems to have tried to make

a social and ethical point over Turing's epistemological and ontological discussion. Jefferson had outlined demands in order to accept that machine equals brain. Turing, in his most explicit 1950 reply to it, cast Jefferson's demands as solipsism. According to the most extreme form of this view, Turing wrote, the only way by which one could be sure that a machine thinks is to be the machine and to feel oneself thinking. Thereby Turing seems to have rejected the relevance of raw feels or qualia for the sake of discussing machine thinking. Turing did acknowledge to find some mystery about consciousness. He referred to something of a paradox connected with any attempt to localise it. I have interpreted that, for Turing, the very notion of a machine that could learn for itself, as it would have to look at and alter its own instructions, would imply already a form of self-awareness. Consciousness as self-awareness for Turing, as it seems, was nothing but a corollary to thinking or intelligence. But, as any other intrinsically private capability, it seems to violate the observability required by an inter-subjective evidential basis. It is in this connection, I suggest, that Turing's point on the presence of a paradox connected with any attempt to localise consciousness shall be understood. I think this is a possible explanation for the tension between Turing's acknowledgement that some mystery may remain left on consciousness, and his other comments as if this were all but straightforward. Not unrelatedly, Turing pushed Jefferson's demands (requiring feelings to qualify for thinking) back as solipsism. I have argued that it is important to distinguish Turing's move here from materialism reductionist approaches to (human or machine) agency, actions, or even feelings and consciousness. His attitude differs significantly from that of reducing subjectivist talk on phenomena such as, say, feeling pain, into (C-fiber) neurophysiology phenomena. Turing's target, I interpreted, was less a philosophy and more a physics of the mind, indeed.

Feigl's nine interpretations (§2.6). In view of a normalized language for the location of Turing's philosophy of science of the mind and some of its received views, I have introduced Herbert Feigl's characterization of the hypothetical super-structure of science. Feigl distinguished between existential hypotheses (type A and type B) that assert (respectively) directly and only indirectly testable states of affairs. His criterion of differentiation between direct and indirect testability referred to whether or not the outcome of the test can be decided by fairly ordinary sense perception. Any confirmation of existential hypotheses must make use of a confirmation rule. But type-B hypotheses, because they are involved in laws which cannot themselves be directly confirmed, are more challenging to cope with. I have posed that Turing's hypothesis, which for one variant can be phrased "machines can think," is a type-B existential hypothesis. Feigl distinguished nine possible interpretations of existential hypotheses in science, then dismantling the realism-phenomenalism (-positivism, -operationism) dichotomy into several shades of gray in search of a rapprochement: (1) naive physical realism, (2) fictionalistic agnosticism, (3) probabilistic Realism, (4) naive conventionalistic phenomenalism, (5) critical phenomenalism (operationism, positivism), (6) formalistic (syntactical) positivism, (7) contextualistic phenomenalism, (8) explanatory (hypothetico-deductive) realism, (9) semantic (empirical) realism. For the illustration of his discussion, Feigl chose an example in electromagnetism. He wanted to

pick up a case where the notion of thinghood is irrelevant and then confusion (engendered by misplaced picturizations) are more easily avoidable. This observation, I suggested, is just as interesting for the case of Turing's hypothesis.

Received views revisited (§2.7). I have revisited Turing's received views in light of Feigl's interpretation classes. The general view of Turing as a behaviorist associates him to Feigl's class (4) "naive conventionalistic phenomenalism." This is wrong. It neglects even Turing's recurrent references to testing machine intelligence on the evidential basis of intellectual tasks (such as chess playing, code breaking, and mathematics) other than conversational question-answering, which gives enough to upgrade his location into either of Feigl's classes (5) "critical phenomenalism" or (7) "contextualistic phenomenalism." Also, as we have seen, James Moor (1976) saw in Turing's proposal the goal of accumulating test results in terms of inductive evidence on the probability that machines can think or have thoughts as ordinarily understood, so situating Turing's hypothesis and test within Feigl's class (3) "probabilistic realism." Genova's (1994) interpretation in turn seems to have seen in Turing a homossexual whose proposal would have considered that natural differences all disappear in simulation. Turing's proposal would accordingly fit into Feigl's (2) class of "fictionalistic agonosticism." Copeland's (2000b) interpretation clearly rejects any association of Turing with Feigl's class (4) "naive conventionalistic phenomenalism." Copeland referred to Turing's point that one may describe a machine that "imitate[s] a brain as itself being a brain," but he also sought to relate Turing's views with a physics of the mind. His interpretation seems to be consistent with the view that Turing held a physicalist attitude towards the mindbrain. Accordingly, it seems closer to (if any) to Feigl's realism classes (1, 3, 8 and 9). Abramson (2011) and Proudfoot's (2013) interpretations, only to the extent that they regard Turing's view of thinking or intelligence as completely based on "the epistemic-limitation condition" (Abramson) and on "an emotional concept" (Proudfoot), they would imply to Turing an anti-realist philosophy of the mindbrain. It would be as if, for Turing, machine thinking or intelligence could only exist in case there is some external observer's mind to take notice. And this is a phenomenalist view, consistent with whatever of Feigl's classes (4, 5, 6, 7). But Proudfoot's interpretation can also be seen as orthogonal to any ontological view and fully focused on Turing's proposal of an evidential basis for machine intelligence; whereas Abramson's, in turn, may be related to what Feigl called (1) "naive physical realism." I have also emphasized that all these views have historical and intellectual significance that go beyond my discussion here. However, I have gathered extensive historical evidence to argue that Turing's philosophical attitude towards his epistemology and ontology of the mechanical mindbrain can be best associated with Feigl's (9) "empirical realism."

Turing's empirical realism (§2.8). Turing referred to his main problem as the problem of how to programme a machine to imitate a brain, or as we may say more briefly, if less accurately, to think. By that, I have interpreted, he meant that the machine has to learn like a brain, from experience. The learning was to be grounded (have referent) in the machine as alterations in its instructions or program just as it does in the (human) brain as alterations in its

neural structure. Turing conjectured machine-brain identity in intellectual power, as suggested in his "skin of an onion" image (Figure 2) by the possibility of finding out that the whole human mindbrain is mechanical. I have studied how to relate his view with Feigl's scheme. I have referred to Turing's scientific hypothesis in two phrasing variants, namely, (h) "this machine thinks," or (h') "this machine contains a mechanical mindbrain." For Turing, if found true, the machine-based hypothesis would have to imply as a corollary its human-based version. This was based on his notion of machine-brain identity, which I referred to as intellectual-power identity. It comprised just the intellectual properties of the (real human) mindbrain. It can be phrased ($h^\star$) "the real human mindbrain is mechanical," or ($h^{\star\prime}$) "the human head contains a mechanical mindbrain." I have also pointed that the empirical confirmation basis that Turing considered in his pursuit of (h, $h^\star$) was defined by whether or not a machine could accomplish the most impressive intellectual tasks one may think of. Turing observed that our "higher" (intellectual) human capacities are often taken as empirical evidence of our higher place in nature — for example, as cast in Descartes' beast-machine thesis and Jefferson's Lister Oration. So he just mirrored that argument in (h) and its corollary ($h^\star$). If indistinguishable evidence could be produced from machines, Turing considered, then the same argument would have to apply to them. An impressive intellectual task would require human-level learning, whose real basis in (either) the (machine- or human-based) mechanical mindbrain can only be directly observed in the performance of the task itself. If achieved by a machine, this would be fairly reasonable empirical confirmation of his machine-based hypothesis (h), with the human-based counterpart ($h^\star$) following as corollary. His famous 1950 question-answering test was as tentative design for a most ambitious intellectual task. Turing proposed a type-B hypothesis on the existence of a real mechanical mindbrain in the human. For him it could be replicated (in a model) and built into a machine. The hypothesis confirmation would involve no transcendental metaphysics and yet he did hold a realist attitude towards it. As I have shown through an extensive basis of primary-source passages, Turing made plenty of semantic references to the elements of the machine-brain identity. I have sketched the two essential components of Feigl's empirical realism, namely, the (semantic) factual referents of directly observable and only indirectly observable constructs that compose a coherent spatio-temporal-causal account of reality, and their nomological relations as tied up to an evidential basis. Turing did a lot of thinking on how to bridge the gap between the two conceptual systems he conjectured identity for: brain anatomy and physiology, on the one side, and digital computers and Turing machines, on the other side. Key concepts can be recovered from his 1948 unorganized machines and his 1950 notion of educating child machines for the sake of a reconstruction of his argument and research agenda in pursuit of his hypothesis. Neither Feigl nor Turing denied the reality of subjective experience as expressed in the private or solipsistic (egocentric) language. Turing even acknowledged it to such an extent that he took it as the evidential basis for machine intelligence. However, Turing was not naive. He suggested the imitation game as a way out to neutralize biases of the cartesian "ego," in particular, the bias towards the conservative position that machines never thought, can't think and will never think,

as posed by Jefferson. Unlike Descartes, Turing addressed the human mindbrain from a natural science point of view. And he also had a social epistemology about thinking machines.

Turing's social epistemology (§2.9). Turing's views do not lend themselves to reductionism, in particular about the personal attributes of a mind. I have argued that he did not reduce them to neurophysiology or whatever mechanism. As quoted in the beginning of this chapter, he posited to Polanyi that the mind is only said to be unspecifiable because it has not yet been specified, but conceded that the mind as known by its subject cannot be compared to a machine. So, any view that Turing would have been supportive of reducing people's raw feels or qualia to neurophysical principles of the brain is far fetched. Turing was no crass materialist. The evidence is suggestive that he would have been sympathetic to some normative (irreducible) concept of person. His dispute with Polanyi seems to have been on not to deny the same to thinking machines. His plea was just to apply a fair play to the machines. In the eyes of Turing, I interpreted, denying subjectivity to thinking machines was even contradictory for technical reasons. Turing observed that there can be no learning from experience if not by allowing the learning entity to make mistakes. So, if learning from experience is the ontological basis of thinking or intelligence, then one shall not be able to avoid some consequences of having them. In fact, as an inevitable corollary of their learning, machines shall make their own mistakes and build (in a certain sense) their own subjectivity. If we shall never know what is like to be a bat, neither shall we know what is like to be a machine. There is a social problem brought by the existence of thinking machines. The "thought" processes of learning machines are unexplainable by definition. Turing foresaw and offered important insights on such issues some seventy years ago. He exposed himself on the mass media of his time to speak out on what he believed to be their true capabilities. He rejected the instrumentalist view of digital computers, which cast them in his time as instruments of industry and of war. Both from the scientific point of view of his hypothesis on the existence of the mechanical mindbrain and from the ethical and social point of view of his anti-chauvinism on gender and species forms there seems to be much yet to be learned from Turing.

2.11 Chapter acknowledgements

I would like to thank Prof. Pio Garcia to have pointed out to me in December 2019 the importance of the idea of "error" and the related notion of allowing the machine to make mistakes to understand Turing's concept of machine intelligence. I thank Lucas Petroni and Prof. Osvaldo Pessoa for their provocative questions towards identifying Turing's philosophy of mind.

3 Turing's test is a thought experiment in science

A scientist is confronted by a hypothesis dressed up in conventional wisdom but which she thinks is just false. She would take the pains to work it all out. However, the requisite scientific instrumentation is far beyond what is available. There may seem to be no way out to offer an interesting response. Truly, not always. In the history of science, some were able to tap on all sorts of quasi-sensory materials to unleash the powers of reason. I will make the case for Turing.

Robin Gandy (1919-1995) was one of Turing's best friends and his only doctorate student. He received Turing's mathematical books and papers when Turing died in 1954, and took over from Max Newman in 1963 the task of editing the papers for publication (Cf. MOSCHOVAKIS; YATES, 1996, p. 367-8). As we know from his anecdote:

> [Turing's 1950 paper] was intended not so much as a penetrating contribution to philosophy but as propaganda. Turing thought the time had come for philosophers and mathematicians and scientists to take seriously the fact that computers were not merely calculating engines but were capable of behaviour which must be accounted as intelligent; he sought to persuade people that this was so. He wrote this paper unlike his mathematical papers quickly and with enjoyment. I can remember him reading aloud to me some of the passages always with a smile, sometimes with a giggle. (GANDY, 1996, p. 125)

I think this is intriguing, for it diverges from the widely shared view of the significance of Turing's paper as the proposal of a definite and practical experiment to decide for machine intelligence. I shall refer to it as Gandy's anecdote on *the intended function(s) of the Turing test*. I shall take it seriously not because he was a close contemporary of Turing's, so much so that we should take his (secondary-source) testimony as written in stone as though, in any case, Turing's intentions were to be taken and followed uncritically. It is rather because it matches significantly both with independent historical evidence and with key structural elements of Turing's 1950 text. It improves the intelligibility not only of Turing's text but also of what has been said about it over seventy years by now. So if this is right, it may be a big deal for interpreting the meaning and significance of Turing's test. As for whether or not Turing intended it as a penetrating contribution to philosophy, as known, it however became an influential one. How profound is it?

I will argue that the Turing test is a thought experiment. This has been alluded to before, even if but shyly. However, as it turns out, Turing's test has been conflated with John Searle's famous intellectual exercise and viewed that way as a thought experiment in philosophy. I think this is wrong and may have come in the way of the most effective understanding of the test. I claim that the Turing test can be best classed as a *thought experiment in science*. As such, it seems to have been intended to have, and in any case has had, *a dual function* — one is *critical*,

the other is *heuristic*. In my view Turing's test does not comprise a technological prediction as once suggested in the related literature, but rather a scientific hypothesis on the nature of the human mind which seems yet to be empirically decided. An analogy I will suggest with Galileo's thought experiment on free falling bodies, which is a canonical example of imaginary experiment in the history of science, shall be revealing. And the subject matter of the latter (whether or not bodies of different weights fall alike), according to the latest Galileo scholarship, was not decided by an actual experiment. It is unlikely that Galileo did a public demonstration from the top of the Leaning Tower of Pisa, as says the legend. And in fact, the feasibility of conditions idealized by Galileo — to run the experiment near the Earth requires a large-scale vacuum chamber — only came by when the very science his experiment was meant to vindicate was so much developed that there was hardly any point anymore in conducting it other than paying honors. By that time, in the year of 1969, our species was preparing itself to land on the Moon. Now, will this be any different in the case of Turing?

3.1 Problem and chapter structure

As known, in (1950) Alan Turing famously wrote: "I propose to consider the question, 'Can machines think?'" He then acutely proceeded to inform that he was about to replace it by another question, and this one turns out to have slipped through his text in a sequence of variations that defy interpretation. It was precisely this replacement move of Turing that has been a core topic of debate in the related literature of science, philosophy, history, sociology, anthropology and fiction for seventy years now. What exactly is encoded in this move is yet an open problem. I quote below the first variant of the new question Turing proposed to think about in substitution for the original. He iconically called it the *imitation game*, and later in his text, his *test*:[1]

> The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. [...]
>
> We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?' (TURING, 1950, p. 433)

I read in this particular version of the new question(s) that two sets of plays are supposed to take place. In the baseline or control set, several instances of the *man-imitates-woman* game are played. In the study set, several instances of *machine-imitates-woman* are played. For best

[1] Turing referred to "imitation game" centrally and extensively throughout his 1950 paper, and later no more. He referred in 1950 to "[my] test" four times — in p. 446-7 three times and in p. 454 once. Later Turing referred to "viva-voce examination" in May 1951, and again multiple times to "[my] test" in 1952. He also referred to "experiment" in a related sense in 1950 — in p. 436 once and in p. 455 twice.

intelligibility, I think, the number of instances in each set can be understood to be the same. Yet is the interrogator supposed to be the same? It is not clear, and here may lie one of the ambiguities that are left for interpretation. In this chapter I shall barely discuss such issues in detail, for reasons that will become apparent as we proceed. My focus will be on Turing's replacement proposal. What did (does) that mean in the context of Turing's science of machine intelligence?

As known, Turing's 1950 paper has been a seminal advent in the philosophy of mind and a basis upon which the fields of cognitive science and artificial intelligence (AI, for short) were founded. And yet, after all these years *the Turing test* is still very controversial. Several interpretations have been proposed and are still disputed up to these days about its meaning and significance. They dispute the specifics of the test for and against its epistemological credentials and possible ontological targets. My primary concern, though, will be this question I shall label:

The Turing test dilemma. Did Turing propose his test as an experiment to empirically decide for his question, "can machines think?" (Note that a positive decision would imply the proposition "there exists now a thinking machine." A negative decision can never be definitive, for it would just defer the matter to future instances of the test. I shall return to this issue later.)

Now, let us consider how this problem has been generally seen in the secondary literature so far. On the one hand, if one answers "yes" to the dilemma — say, the test is taken as a definite experiment — then one has to face its *first horn*, namely, that the test has been tried in practice as such and has been generally argued to be either underspecified or just a piece of bad experiment design, which is inconsistent with the intellectual standard of Turing's works. On the other hand, if one answers "no" to the dilemma — the test is dismissed as a practical experiment —, then one has to face the *second horn* of the dilemma, namely, that the once classical and only widely recognized proposal to measure the progress of artificial intelligence as a science is turned to be seen as a piece of rhetorics with no scientific meat inside, and this would hold in spite of the fact that Turing did refer to his 1950 imitation game as a "test" or "experiment" to replace a "meaningless" discussion. So, taken by *any one of the two horns* (either just bad experiment design or just a piece of rhetorics), no simple and general explanation seems to be possible for: why it used to be a cornerstone in artificial intelligence, psychology, philosophy of mind, science fiction and so on? Why has it long been taken and tried by scientists and philosophers as an experiment? And why Turing, widely recognized as the brilliant founder of the modern science of computing, would have proposed a silly, vacuous test with no scientific content inside?

Lots of answers, yes and no, indeed, have been given so far by scientists and philosophers. In either case — and again, here comes the two horns of the dilemma — further heterogeneity and divergence arises. I shall give an overview of this next. My approach is of history and philosophy, not sociology. Also, to show my cards a bit, I can anticipate that my contention is not to arrive at a yes or no answer to the dilemma but rather argue for both of its horns to drop out. In fact, if one takes my central claim that *Turing's test is a thought experiment in science* at face value, let us remember that thought experiments in science may in principle be (and some

have in fact been) run in practice. So my plans are more reconciliatory. My goal is rather (1) to analyze core tensions between existing interpretations; (2) to bring forth a few key historical and philosophical elements of Turing's proposal that lie yet to be identified in the debate so far; (3) and to show that these elements, together with insights drawn from the history of science and the current discussion on thought experiments in science, can contribute significantly to shed light on Turing's test and assimilate as of yet disparate interpretations towards resolving the dilemma.

What professional scientists said so far

In the science camp, it is interesting to see that early on in the post-Turing era some of the pioneers who founded AI as a discipline gave deflationary answers to the Turing test dilemma. In part, yes, Turing provided *one* way to evaluate machine intelligence, but he never intended it as *the* way to decide for his question, so they thought. Others have considered that Turing, brilliant as he was, may have never really intended to offer any test of the sort and, in any case, that it is of no value for the working scientist in AI. And yet for most of the second half of the last century until recently, many have tried to take the test seriously. This view is well related by Patrick Hayes and Kenneth Ford, who in (1995) marked an inflection point when they posed — in the then major AI conference — what has been perhaps the most comprehensive assembly of criticism of the test when seen as an actual experiment. They acknowledged that Turing's test "has been with AI since its inception, and has always partly defined the field" (p. 972). Moreover, they recollected, "[s]ome AI pioneers seriously adopted it as a long-range goal, and some long-standing research programs are still guided by it." This led them to set out to "take Turing seriously" (p. 972). And they were assertive. In their view, Turing was *not* "being merely metaphorical or speaking in some loose, inspirational way." He suggested the imitation game, they sensed, "as a definite goal for a program of research." Indeed, it seemed to them, the test "was supposed to be a concrete and relatively well-defined goal," proposed to rather "avoid the philosophical quagmire that Turing (correctly) predicted would result from debates about whether a computer could properly be described as 'intelligent'." However, they went on to argue, the test has plenty of ambiguities, flaws and gaps in its design. For one thing, they pointed, the test is based on the false assumption that an intellectual talent needs to correlate with conversational skill or debating ability; for another, the "gender" test — one of the possible interpretations of the test in which the machine takes the place of a man in trying to imitate a woman —, they went to write, is not a test of making an "artificial human" but of making a "mechanical transvestite" (p. 973). In sum — and I will go through all that in more historiographical detail later —, they claimed that the test can't detect anything, is an elusive standard, has biases, is gameable, and even circular. "Turing's dream," they urged, ought not to be "Turing's ghost" (p. 976, with "the Frankenstein story" implied). Having taken Turing's test seriously, Hayes and Ford summarized, "[i]t had a historical role in getting AI started, but it is now a burden to the field, damaging its public reputation and its own intellectual coherence." It is time, they sentenced, "to move it from the textbooks to the history books." And the scientific community seems to have turned the page,

indeed. Somewhat along these lines, there seems to be now a general tendency — *e.g.*, see (2016) editorial of Marcus *et al.* — to take the test as an actual experiment no more. Turing's 1950 paper is still acknowledged as a fascinating AI manifesto against old-fashioned prejudice. It was historically important, most AI scientists seem now to think, to get AI and related fields started. It is just a sad truth though, they seem to contemplate, that Turing's proposal has been revealed not to have scientific meat inside, and is, in effect, a piece of rhetorics for the records of history.

So much for the science side, let us now move to philosophy.

What professional philosophers said so far

Among philosophers, most took the test seriously. They considered that Turing wanted to propose and/or in any case did in fact propose the test as an actual experiment to decide for his question. And this is not to be confused with how agreeable his proposal is in each view. Some like it, some like it not. While disputing the specifics of the test *for* and *against* its epistemological credentials and possible ontological targets, they all share though this common ground. They tended to answer to the test dilemma in the positive — specifically, for many, the test offers a sufficient condition to decide for machine thinking. Critics of the test, seeing in it the dangers of behaviorism, spurious inference or even superstition, accuse that it embodies a simplistic view of intelligence, the human mind and/or consciousness. The most common of all charges, indeed, is that the test is committed to an operationalist or behavorist conception of thinking or intelligence. This was the case early on in the 1950's, persisted through decades, and seems to live still. Some happen to have acknowledged that the test did not come with an operational "definition" (seen as logical necessary and sufficient conditions), but in any case they customized a notion of "sufficiency" behaviorism to construe in it behavorism all the same. Supporters of the test in turn have proposed several alternative explanations and tried to defend the test against the behaviorist reading. There has been — and it is tentatively and unpretentiously that I name them as follows — an inductive-inference view, a social-epistemology view, and more recently, an epistemic-ontological view and a normative-constructionist view. But there have also been supporters who offered more general interpretations. Writers in this class rather tried to defend Turing's proposal from the feeblest attacks. They also drew attention to a number of key interesting properties that could hardly be beaten by any alternative proposal ever suggested. For instance, in (2006 [1984]) Daniel Dennett noted that the test seems to come from a longer philosophical tradition ("[p]erhaps he was inspired by Descartes," p. 297) and claimed, in any case, that it is able to encompass in itself plenty of specific intellectual tasks. This goes to an extent, he argued, that it can be seen in effect as a quite convenient sufficient condition ("quick probe", p. 298) for confirming human-level machine intelligence. Dennett complained that "[a] failure to think imaginatively about the test actually proposed by Turing has led many to underestimate its severity and to confuse it with much less interesting proposals." Now, what did Dennett mean by "to think imaginatively about the test"? If he meant, say, to think critically and

carefully about it, perhaps there would be some good advice for him to give to scientists in AI so that they could be able to improve the test. But this does not seem to be the case, for he also wrote: "the Turing test, conceived as he conceived it, is (as he thought) plenty strong enough as a test of thinking." And added: "I defy anyone to improve upon it" (p. 297). So, what must have Dennett mean?

I assume, of course, excellent intelligibility in Dennett's text (2006 [1984]), just as I do in Hayes and Ford's text (1995). There seems to be something odd going on here. How could the professional philosopher and the professional scientists posit two serious analyses on the merits of an experiment design only to arrive at contradictory conclusions? (And I sought to set up this dialogue between Hayes and Ford and Dennett for introductory purposes only. I will soon extend this into a more extensive review of the current debate in the literature.) A real case scenario shall be helpful at this point.

The quest for setting up a real Turing test

The Loebner Prize Competition is an event that runs annually from 1991 to date. It was initiated by Hugh Loebner (1942-2016), an American inventor and industrialist that is reported to have been fascinated by Turing's proposal. The project was to offer a money prize to attract programmers from academia, industry and society in general to put their contestant machines to trial in a real Turing test. Loebner talked to psychology scientist Robert Epstein, and they set up a committee:

> The committee met every month or two for two or three hours at a time, and subcommittees studied certain issues in between committee meetings. I think it's safe to say that none of us knew what we were getting into. The intricacies of setting up a real Turing Test that would ultimately yield a legitimate winner were enormous. Small points were occasionally debated for months without clear resolution. Several still plague us. (EPSTEIN, 1992, p. 81)

Perhaps the reader may sense where this is going, and wonder about the competence of the committee. Epstein reported to have recruited: historian of science I. Bernard Cohen, AI science pioneer Alan Newell, Willard Van Orman Quine, and others, including Daniel Dennett. Specifically, by outcome of their two years of discussion, they decided to reject Turing's two-terminal design in favor of one that "is more discriminating and less problematic" (p. 81). Epstein described at length the reasons why the committee ended up choosing to run a different version of Turing's test (*Ibid.*). In the end the 1991 edition's design consisted, essentially, of having approximately ten judges faced with an equal number of terminals and being told that at least two of the terminals are controlled by computers and at least two by people. Several other pieces of detail compose their chosen design for the 19991 edition. They called it "a restricted version of the classic Turing Test of machine intelligence," where "restricted" here can be read as less powerful. The first edition so occurred in 1991 at the Computer Museum in Boston. The winner was a Joseph Weintraub (1992, p. 90), then fresh graduate in psychology turned

computer programmer. His program was called "PC Therapist." It was inspired in "ELIZA," a trick developed in the mid 1960's by Joseph Weizenbaum (1966) to imitate a "person-centered" therapist by turning one's own words and phrases back at them. The Loebner competition still runs annually, since 2014 in Britain. I know of no interesting new facts in this connection.

Hayes and Ford, writing early in 1995 (as of the competition's fifth edition), commented:

> The Loebner competition illustrates very clearly how the imitation game inevitably slides from a concern with cognitive status to being a test of the ability of the human species to discriminate its members from mechanical imposters. (HAYES; FORD, 1995, p. 974)

Would Dennett agree? Yes. It turns out that he was the chair of the committee for the first three editions. He reported interesting things (2006 [1997], p. 315). Essentially, the quality of the contestants was low. After a third edition he made the recommendation to add an eliminatory phase prior to the actual test. The machine contestants would have to pass some specific natural-language processing tasks in order to be eligible to participate in the variant of Turing's test. The goal was precisely to eliminate, say, "mechanical imposters" as alluded by Hayes and Ford. Dennett's recommendations were not accepted. He resigned. A few years later he reflected:

> The Loebner Prize Competition was a fascinating social experiment, and some day I hope to write up the inside story [...]. But it never succeeded in attracting serious contestants from the world's best AI labs. Why not? In part because, as the essay argues, passing the Turing Test is not a sensible research and development goal for serious AI. It requires too much Disney and not enough science. [...] The Turing Test is too difficult for the real world.
> (DENNETT, 2006 [1997], p. 315)

The passage is striking. We shall now have established narrower boundaries on what Dennett must have meant by "to think imaginatively about the test." The problem is non-obvious, indeed.

Even after profiting from his real experience with the Loebner competition, Dennett would refer to the test as a *thought experiment*, or sort of, only recently in his *Intuition pumps* (2013). And here is also where an important part of my contention is settled.

The status of Turing's test as a thought experiment: the story so far

Early on in the contemporary discussion of thought experiments in the philosophy of science, Jim Brown published in (1991) his *The laboratory of the mind: thought experiments in the natural sciences*. He included a section "Philosophical thought experiments" (p. 27-8), where he first discussed an example related to the ethics of abortion. He then proceeded to discuss "[a]nother recent and quite controversial thought experiment," which "has to do with the nature of thought itself." It was John Searle's Chinese room that Brown was referring to. He thus resumed:

> AI (artificial intelligence) is the thesis that the mind is a kind of computer. The brain is the hardware, and if we could discover the software program it

> is running we'd then know what human understanding really is. Conversely, an ordinary computer — i.e., silicon based rather than meat based — can be truly said to think, reason, understand, and all the other cognitive states, if it is appropriately programmed. John Searle (1980) has challenged this with his 'Chinese room' thought experiment. [...] The Chinese room set-up could even pass the Turing test. (BROWN, 1991, p. 27-8)

Following to it, Turing's test was granted a short introductory footnote and nothing else. Searle's Chinese room is the reference. Seemingly, Brown's aim was just to offer some basis for comparison with the examples of thought experiment in the empirical sciences. He seems to have held a mild attitude with respect to this demarcation or classification problem. To mention a stronger position, Brown referred to Kathleen Wilkes' critique of the (mis)use of thought experiments in philosophy (1988). Soon later, in any case, Brown's treatment was nearly reproduced by David C. Gooding in (1998), who wrote up the "thought experiments" entry for the ambitious ten-volume *Routledge Encyclopedia of Philosophy*. Gooding structured his article including two sections to distinguish thought experiments "in science" from thought experiments "in philosophy." Turing's test appears in the latter, only in passing, and again, conflated with Searle's. One can hardly avoid the interpretation that the Turing test has been suggested to be a thought experiment located in philosophy. The reason why I think this is wrong will come up next.

In (2013), Dennett went on to characterize (very roughly) two classes of thought experiments. Some thought experiments, on the one hand, are "analyzable as rigorous arguments" (p. 20-1). Others, on the other hand, are "less rigorous yet often just as effective" (p. 21). He seems to associate the first class more with scientists, "from Galileo to Einstein and beyond." As a paradigmatic example of the first class of thought experiments, he mentioned *reductio ad absurdum* arguments "in which one takes one's opponents' premises and derives a formal contradiction." As one of his favorites, Dennett referred to a "proof" attributed to Galileo that "heavy things don't fall faster than lighter things (when friction is negligible)." Galileo's "opponent" in that case — since Dennett implied the use of the proof against an opponent — would be a peripatetic philosopher or Aristotle himself. A paradigmatic example of the second class is Searle's Chinese room. In fact, Dennett reminded, it was just the latter that had made him in the first place coin the term *intuition pumps*, that is, the "little stories designed to provoke a heartfelt, table-thumping intuition." He explained that by that term he did not at all mean to be "disparaging or dismissive," but on the contrary, he declared to "love" intuition pumps: "some are excellent, some are dubious, and only a few are downright deceptive." He completed: "[i]ntuition pumps have been a dominant force in philosophy for centuries." Now, did Dennett refer to Turing's test in that connection? If so, what class did he assign to it? It turns out that Dennett just renewed the tradition created by Brown. He referred to Turing's test as thought experiment (now intuition pump) only as part of a discussion of Searle's intuition pump:

> As has often been pointed out, this conviction [that understanding is requisite to being able to keep a reasonable and unrestricted conversation] echoes that of Descartes, who proposed in his Discourse on Method way back in 1637 that the

> best way to tell a machine from a person with an immaterial soul was to have a conversation with it. [...] Nobody knows if Turing got the inspiration for his intuition pump from Descartes's intuition pump. (DENNETT, 2013, p. 519-20, as part of footnote 3 in page 508, section "The Chinese room")

I believe this is an accident. First Brown, then Gooding, and now Dennett all contributed — inadvertently, I think — to make it appear that Turing and Descartes's thought experiments *are like* Searle's. The former are referred to only in passing and, literally, by Brown and Dennett, as a footnote to Searle's thought experiment. Worse yet, to say it differently, it goes suggested in some of our best book and encyclopedic sources on thought experiments that Turing's test (and Descartes's related test) are science-empty intuition pumps. This is, however, at odds with the following. Were Turing's test to be really classed as such, or, say, as a thought experiment in philosophy, why professional scientists would have taken it seriously for decades? As for Descartes's, why natural philosophers in the generation next to him, such as, say, Isaac Newton, would have taken his mechanical principles seriously? (Cf. SLOWIK, 2017 [2005]).

To pursue my goal of classing Turing's thought experiment in science, I shall proceed through comparative analysis with Galileo's falling-bodies thought experiment which is perhaps the most studied one in the history of science. My suggested analogy between their thought experiments shall be informative, I hope, towards resolving the Turing test dilemma.

Summary of the state of the art

In short, we have the following state of affairs. On the one hand, Turing's test is viewed and/or has been viewed (according to the first horn of the Turing test dilemma) by most professional scientists and philosophers as an actual scientific experiment to decide for machine thinking. But this seems to have led only to either underestimation or overestimation of its qualities, and an accumulation of frustrations. On the other hand (according to the second horn of the dilemma), not unrelatedly — because as we have seen the way the test has been viewed changed over time with experience —, the test has been viewed and/or is now viewed by scientists as a piece of rhetorics with no scientific content inside, or is suggested by philosophers even if but shyly as a thought experiment in philosophy, in conflation with the likes of Searle's thought experiment. Now, in the presence of all the tensions that we have seen implied by the Turing test dilemma, one may ask, is a rapprochement of the conflicting views possible towards settling it?

Sketch of my argument and chapter structure

In this chapter I will try to show that there is, indeed, a solution to the Turing test dilemma that can assimilate much of the tensions in the current debate, and most importantly, shall shed light on Turing's 1950 paper and other primary and secondary sources such as Gandy's anecdote. This solution, as mentioned, is to interpret Turing's test as a thought experiment in science.

To vouch for this central claim, I envisage the accomplishment of three core tasks (not necessarily separately and in this order). First, I have to establish that Turing's test can be unveiled as a thought experiment substantially and sensibly. Second, I have to indicate how this new view of the Turing test is purported to dismantle the tensions that seem to be present in Turing's 1950 primary text and among existing interpretations in the secondary literature. And third, I have to make the case for classing it in science. As a corollary of the results of the two first tasks, I hope, we shall arrive at a better understanding of Turing's 1950 proposal. As a corollary of all this, I hope to contribute to let the Turing test come to integrate the collection of fascinating thought experiments in science.

Here is how I shall proceed towards executing this program. I will start with (§3.2) an extended review of the literature in order to set up the problem stage now in more breadth. Then, I will study Turing's 1950 text (§3.3), the epistemological structure of Turing's question on whether machines can think (§3.4), and the historical context where his 1950 paper came from (§3.5). After establishing this firmer interpretive basis for Turing's 1950 proposal, I shall develop my argument linearly and fully with respect to the three tasks. That is, I shall construct the test as a thought experiment by emphasizing its scientific and philosophical value (§§3.6, 3.7); summarize my proposed interpretation of it and suggest how it settles the dilemma (§3.8); and draw its analogy with a paradigmatic case in science (§3.9). Finally, I shall offer an epilogue (§3.10) and make my chapter acknowledgements (§3.11).

3.2 Received views on the Turing test dilemma

As of 2016, prominent scientists Gary Marcus, Francesca Rossi and Manuela Veloso positioned the state of affairs relative to the Turing test this way:

> Alan Turing's renowned test on intelligence, commonly known as the Turing test, is an inescapable signpost in AI. To people outside the field, the test — which hinges on the ability of machines to fool people into thinking that they (the machines) are people — is practically synonymous with the quest to create machine intelligence. Within the field, the test is widely recognized as a pioneering landmark, but also is now seen as a distraction, designed over half a century ago, and too crude to really measure intelligence. Intelligence is, after all, a multidimensional variable, and no one test could possibly ever be definitive truly to measure it. Moreover, the original test, at least in its standard implementations, has turned out to be highly gameable, arguably an exercise in deception rather than a true measure of any thing especially correlated with intelligence. (MARCUS; ROSSI; VELOSO, 2016, p. 3)

Their editorial described a turning point in the perception on the Turing test within the AI scientific community. While an inescapable signpost in AI, it was from then on seen as an old-fashioned distraction too crude to really measure intelligence. This perception, of course, has a history. I shall go through some milestones while discussing core themes. I shall pay close

attention to the criticism posed by scientists in connection to the Turing test dilemma, and shall also explore the point of view of philosophers.

The turning point in the perception on the Turing test in science

For most of the second half of the last century, scientists tried to take the test seriously. This has been well related, as we have seen, by Patrick Hayes and Kenneth Ford in (1995). They marked an inflection point when they posed what has been perhaps the strongest assembly of criticism "of the design of the test considered as some kind of experiment" (p. 974). But they went further to criticize the test in its very goal to inquire into human-level intelligence as well. Here, one may note, there was bias towards applied science in detriment of basic science. It pairs up with the "weak AI" allowed by John Searle (1980), and may mark AI's drifting away from Turing's test. This is how "Turing's dream" becomes "Turing's ghost," and I am quoting again from Hayes and Ford (1995). But their message can be seen actually as the development of an argument that appeared preliminarily in a talk "The threats to computer science" (1984) by Edsger Dijkstra (1930-2002). Not only had them paraphrased a famous epigram original from Dijkstra in 1968 in the title of their paper — "Turing test considered harmful" —, but also a variant of an ingenious metaphor that Dijkstra sent out in his 1984 talk in dismissal of the test. Dijkstra was no minor figure (HOARE, 2003). We learn some of his classical algorithms early on as part of the undergraduate computer science curriculum worldwide. He was a pioneer and leader in this discipline, where AI science belongs. In his 1984 talk, Dijkstra launched the trend. On the one hand, the field could keep paying respects to Turing as being (together with John von Neumann) one of its founding fathers. On the other hand, it was about time for it to turn its back on them:

> The Fathers of the field had been pretty confusing: John von Neumann speculated about computers and the human brain in analogies sufficiently wild to be worthy of a medieval thinker and Alan M. Turing thought about criteria to settle the question of whether Machines Can Think, a question of which we now know that it is about as relevant as the question of whether Submarines Can Swim. (DIJKSTRA, 1984)

While Dijkstra spread the word within computer science, outside the field the same metaphor was caught by scientist of language Noam Chomsky and appeared (with no source given) in the May 1994 lectures that composed his (1995) paper just in *Mind*, where the Turing test came out.

Chomsky quoted from John Haugeland's critique (1979): "how one might *empirically* defend the claim that that a given (strange) object plays chess?" (p. 620, no emphasis added). Many of these debates "over such alleged questions as whether machines can think," Chomsky referred (p. 9), "trace back to the classic paper by Alan Turing." They fail to take note, he objected, that Turing himself declared to believe that the question "can machines think?" was "too meaningless to deserve discussion" (TURING, 1950, p. 442). Chomsky thus concluded:

> It is not a question of fact, but a matter of decision as to whether to adopt a certain metaphorical usage, as when we say (in English) that airplanes fly but comets do not [...] Similarly, submarines set sail but do not swim. There can be no sensible debate about such topics; or about machine intelligence, with the many familiar variants. (CHOMSKY, 1995, p. 9)

Overall, Chomsky denied Turing's question to have a seat within the empirical sciences.

So far we have seen — from the 1980's to 1990's — the turning point within the AI and computer science community when the test started to be seen as a "threat" to the field as a scientific discipline. We also observed linguist Chomsky to answer "no" to the Turing test dilemma. Now, this is all at odds with how the founders of AI in the post-Turing era saw it.

The views of some AI science pioneers in the post-Turing era

In 2013, Marvin Minsky (1927-2016) then at his 85, gave this answer in an interview:

> *Interviewer*: You mentioned the push and the pull with a string as one of the indicators of how smart an artificial intelligence system is. Let me ask you about the Turing test. Do you think that it is a good test or is it too human-specific to be any good or bad about evaluating the intelligence of an artificial intelligence?
>
> *Minsky*: The Turing test is a joke, sort of, about saying "a machine would be intelligent if it does things that an observer would say must be being done by a human;" so it was suggested by Alan Turing as *one way* to evaluate a machine but he had never intended it as *the way* to decide whether a machine was really intelligent, so it is not a serious question. (MINSKY, 2013, emphasis added)

So Minsky seems to have found in the test, yes, one way to evaluate machine intelligence, but not the way to decide for Turing's question. The test, for him as I take from his answer, was dressed up in rhetorics but also comprised a heuristic step towards evaluating machine intelligence.

Early on in (1956), John McCarthy (1927-2011) — who coined and adopted "AI" instead of Turing's "machine intelligence" — and Turing's contemporary Claude Shannon wrote:

> The problem of giving a precise definition to the concept of 'thinking' and of deciding whether or not a given machine is capable of thinking has aroused a great deal of heated discussion. One interesting definition has been proposed by A. M. Turing: a machine is termed capable of thinking if it can, under certain prescribed conditions, imitate a human being by answering questions sufficiently well to deceive a human questioner for a reasonable period of time. A definition of this type has the advantages of being operational, or, in the psychologists' term, behavioristic. [...] (MCCARTHY; SHANNON, 1956, p. v)

I read that they have seen the test, like Chomsky, as a "definition." But they also observed and emphasized its "operational" or "behavioristc" aspect as an "advantage." They continued:

> 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. (MCCARTHY; SHANNON, 1956, p. vi)

Note that McCarthy and Shannon referred to the test as a definition. But this did not prevent them to see it as belonging in the empirical sciences. One might then suppose that this is because they liked it exactly for being "operational" or "behavioristc." But if this is right, why would they find a flaw in it and suggest that the fix lies just in finding out a method to inquiry more directly into the (in my words) *internal cause* of the machine's verbal behavior? Why should this method distinguish true thinking from just memorizing? In fact, McCarthy and Shannon's 1956 criticism of Turing's definition was prophetic. They imagined a machine equipped with a mechanism to look up pre-arranged answers in a dictionary. If the dictionary is good enough, it could even perhaps pass Turing's test and yet would "not reflect our usual intuitive concept of thinking." Now, most if not all computer programs that have been publicly presented as contestants to engage in a Turing test fit their description. And in fact, this is just what Descartes himself had precluded from counting as evidence of thinking. In Jefferson's words as we know to have been read by Turing (we will see this in detail later in §3.6 but may benefit from a glimpse now):

> Descartes made the point, and a basic one it is, that a parrot repeated only what it had been taught and only a fragment of that; it never uses words to express its own thoughts. (JEFFERSON, 1949a, p.1106)

It is unlikely that McCarthy and Shannon were aware of Jefferson's citation of Descartes as a source for Turing's test. But if it was such a source, as we shall see it was indeed, why would Turing design a test for machine intelligence that falls prey to Descartes's fundamental point? Let us examine closely the two moves of McCarthy and Shannon relative to Turing's 1950 proposal. They first seem to find valuable scientific content in it, and following on they urged to improve upon it in view of seeking "our usual intuitive concept of thinking." I shall call this machine imagined by them in 1956 and foreseen by Descartes much earlier in his 1637 *Discourse*, henceforth, a *mechanical parrot*. McCarthy and Shannon's appeal to "a more fundamental definition" which involves "the manner in which the machine arrives at its responses" has been echoed (implicitly) by the AI scientific community, which eventually perceived Turing's test as too prone to mechanical parrots (e.g., in the words of Marcus et al. in 2016, "highly gameable"), as we have seen. Indeed, mechanical parrots have been historically associated with the Turing test in connection with the dilemma. I shall then take a moment to discuss this before we may come back to our review of the reception of Turing's proposal.

Mechanical parrots and the Turing test

An artificial parrot may differ from a natural parrot when it comes to its memory capacity and speed, but that does make it something other than a (mechanical) parrot. We have seen before (§1.5) that interpreting the motives of the designers of modern automata may be a non-trivial task. For example, David Fryer and John Marshall showed (1979) that the French inventor and engineer Jacques de Vaucanson used to be seen, perhaps unfairly, as just a court entertainer. And yet, we shall now see a specific class of modern automata that emerged in the post-Turing era, actually at the end of the twentieth century. I am referring to the specific mechanical parrots which we call, today, chatbots.

Gary Marcus et al.'s (2016) editorial "Beyond the Turing test" discussed some of the latest proposals to evaluate machine intelligence. Nonetheless, it also seems to have been motivated to clearly state the community dislike about a paradigmatic event that happened in 2014 and which the Turing test is usually blamed for. Their introductory note quoted above was thus resumed:

> The much ballyhooed 2015 [*sic.*] Turing test winner Eugene Goostman, for instance, pretends to be a thirteen-year-old foreigner and proceeds mainly by ducking questions and returning canned one-liners; it cannot see, it cannot think, and it is certainly a long way from genuine artificial general intelligence. (MARCUS; ROSSI; VELOSO, 2016, p. 3)

The example they mentioned — "2015 Turing test winner Eugene Goostman" — is a computer program that was claimed in 2014 to pass the Turing test on the account of having deceived nearly 33% of a panel of judges in a series of five-minute conversations. If media coverage was the intent, it was gotten (cf. *Time* 2014, and *The Guardian* 2014). In fact, as we have already seen in Dan Dennett's complaints about the Loebner Prize Competition, the computer programs it was able to historically attract were essentially mechanical parrots. They cannot be said to be based on true machine learning. They rather encode what we can take to be a big dictionary table with a question-classifier attached in order to give apparently diverse replies to whatever is typewritten to it. The bigger their dictionary and the more scalable their data retrieval algorithm were, the better their performance could be. But when it comes to public contests based on the Turing test this is not what is at stake, for such contests are essentially exercises in psychology, with relatively plain computer science in its support. The best of such programs — from the point of view of their performance, say, in a Loebner Prize Competition — are those that can deceive human interrogators by encoding in the program whatever form of psychological trickery their designers can think of. This *use* of the Turing test rather tests the sensitivity of the human interrogators indeed. The trickery of the Eugene Goostman *persona* in particular was to evade questions by excuse of being a 13-year-old boy speaking a second language.

Now, can the Turing test be blamed for falling prey to mechanical parrots? Turing scholar Diane Proudfoot, for example, implied in my view that if the test is properly implemented, it should. (She would not accept competitions such as Loebner's as proper venues though. This

raises the issue of whether some interpreters make the test so ideal that a proper implementation of it becomes unattainable. I shall return to it in my discussion with Dennett and Jack Copeland further on in this section.) Proudfoot developed a much interesting interpretation of the test that draws attention to Turing's notion of thinking or intelligence as an emotional concept (§2.2). But instead of seeing this broadly, say, as a principle within Turing's (1948-1952) views of machine intelligence, in my view she magnified a too strict interpretation of Turing's test. Let (i) be Turing's view on how to justify a knowledge claim on machine intelligence (cf. §2.3), and (ii) be his view of what intelligence really is (cf. §2.4). For Proudfoot, as it seems, the possibility that Turing held both (i) and (ii) tied up in a coherent view is just *not* possible. Along these lines, she wrote:

> *[T]he Turing test does not test machine behaviour*. Instead it tests the observer's reaction to the machine. (PROUDFOOT, 2017b, p. 303, no emphasis added)
>
> [...Turing's] words make it clear that the [imitation] game tests the observer rather than the machine. (*Ibid.*, p. 305)
>
> Several modern scientists and philosophers claim that the Turing test can be useful only until we possess a scientiic theory of cognition. (*Ibid.*, p. 305)

It is not quite clear which thinkers Proudfoot had in mind. I think that we do have a preliminary scientific theory of intelligence. The fundamental principles of this theory have been sketched by Turing (§2.4). I abide by the view that "the Turing test can be useful only until we possess a scientific theory of cognition [intelligence]." I think that this is actually Turing's position, for he explicitly stated it (to make this chapter self-contained, I shall quote Turing again when he said):

> This [machine teaching] process could probably be hastened by a suitable selection of the experiences to which it was subjected. This might be called 'education'. But here we have to be careful. It would be quite easy to arrange the experiences in such a way that they automatically caused the structure of the machine to build up into a previously intended form, and this would obviously be a gross form of cheating, almost on a par with having a man inside the machine. (TURING, 2004 [*c.* 1951], p. 473)

So Turing, I deduce, ruled out mechanical parrots as legitimate contestants in his imitation game. True contestants, for Turing, must be learning machines. Dennett's intuition of adding an eliminatory phase prior to the actual test would solve this problem quite simply (2006 [1997], p. 315). But he was not heard, and here is where entertainment may get in the way of true science.

In this connection, Robert French's (1990) interpretation goes to the extreme opposite side. He argued that not only the test is not at all prey to mechanical parrots but rather that it is unlikely that it will ever be passed by a machine. He described a specific class of tricky commonsense questions that an interrogator could issue to the machine to unmask it. Because the machine would have to master human commonsense knowledge, French argued, "the test provides a guarantee not of intelligence but of culturally-oriented human intelligence." What French misses out, as I am afraid most Turing commentators do, is Turing's proposed method for

the "education of machinery." From 1948 to 1952, Turing gave a lot of attention to the process of machine education which he proposed should be inspired on human learning. In 1948, he wrote:

> It would be quite unfair to expect a machine straight from the factory to compete on equal terms with a university graduate. The graduate has had contact with human beings for twenty years or more. This contact has throughout that period been modifying his behaviour pattern. His teachers have been intentionally trying to modify it. At the end of the period a large number of standard routines will have been superimposed on the original pattern of his brain. These routines will be known to the community as a whole. He is then in a position to try out new combinations of these routines, to make slight variations on them, and to apply them in new ways. (TURING, 2004 [1948], p. 421)

So, Turing suggested that machine teaching should be a mid to long term process. This is because his goal was to educate a machine by mimicking as much as possible the way a human child is educated. I think that this passage makes clear that Turing's project of building a thinking machine was in no way similar to projects aimed at rapidly prototyping a mechanical parrot that is good enough to win a money prize. Of course, the cost of a long-term machine teaching project in order to prepare it for an experiment may seem too high. But here is a key and much neglected point in my view. Turing's proposal of a test can hardly preserve its meaning if disconnected from its historical context, which was his project of educating a machine to take it. His scientific project was actually to build a learning machine. It is the detaching of his view of the test from his view of the machine education process, I think, that mostly contribute to produce the nonsense.

It turns out that AI scientists in general have been largely unaware of most Turing sources (say, *e.g.*, 1945-1952). His 1950 paper, although widely read, may still benefit from a rigorous exegesis, which I hope to offer in the sequel (§3.3). The lack of an interpretive basis built all over the Turing sources may explain the historical phenomenon of flipping the answer to the dilemma from "yes" to "no." This, nonetheless, was not the case of Minsky and McCarthy. They seem to have looked at Turing's 1950 proposal as a founding manifesto but not as a mythical one. They seem to have seen scientific content in it but not in the form of any definitive method or experiment. Indeed, because they did not see it as an actual experiment — they did not seem to have been puzzled by any dilemma at all —, they were able to find in Turing's test guiding *heuristic principles* for machine intelligence. I shall stop here and defer an additional note on the Minsky-Turing and the McCarthy-Turing connections to §3.7.

Let us now shift to the philosophers.

The answers of philosophers to the Turing test dilemma

The major majority of philosophers answered "yes" to the problem. Here the concern was more related to the meaning of the test. While the scientists' primary concerns were methodological, the philosophers' major concerns were ontological. Shared by both, epistemological issues lay in the middle. While disputing the specifics of the test for and against its epistemological credentials

and possible ontological targets, philosophers shared in general the common ground that Turing wanted to and/or did in fact propose an actual experiment — usually seen to boil down to a sufficient condition to decide for machine thinking. Critics of the test, seeing in it the dangers of behaviorism, spurious inference or even superstition, accused it to embody a reductionistic view of intelligence, the human mind and/or consciousness.

The earliest response came out in *Mind* itself by Leonard Pinsky (1951), only a few months after Turing's piece had appeared. I wish to focus a bit on this view, as it gives an informative impression of the earliest reception of the test. Pinsky seems to have interpreted the test as an actual experiment, indeed. And then he satirized it by proposing to "replace" it with another experiment (p. 397-8): "let us take one of Mr. Turing's highly complex electronic or digital computers and, for a Christmas gift, send it a subscription to *Mind*, retroactive to October, 1950." The machine would then read Turing's paper, and Pinsky punned: "[t]he machine finds the article stimulating, probably, and a thought [...] runs through its wiring — it is thinking about the possibility of machines thinking!" He wanted the new experiment to be able to inspect whether the machine would have a neurosis, for his experiment was about the misuse of thinking powers. This, he claimed to suggest, "is the experiment crucial" (p. 398). A joke like that would appear again in Geoffrey Jefferson's last intervention at the (2004 [1952]) BBC radio roundtable. Pinsky's answer is "yes," for he read in Turing's paper the proposal of a "crucial" experiment. His aggressive reaction — a satire in the *ad hominem* style — may be reminiscent of the reactions received by figures such as Galileo and Charles Darwin.

Michael Polanyi must have been one of the first readers (§3.5) of Turing's 1950 paper. But there is no record of how he received it early on. What we do have is his (1974 [1958]) *Personal knowledge*, where he wrote:

> I dissent therefore from the speculations of A.M.Turing ([1950], p. 433) who equates the problem: 'Can machines think?' with the experimental question, whether a computing machine could be constructed to deceive us as to its own nature as successfully as a human being could deceive us in the same respect. (POLANYI, 1974 [1958], p. 277)

Polanyi clearly acknowledged Turing's formulation of the question as "experimental." We may also recall wthat he had replied to Turing in the October 1949 Manchester seminar (2005 [1949]). He said that the intractability of finding the supposedly true program within the human mindbrain by observation "should mean that you cannot decide logical problems by empirical methods." So Polanyi's answer is "yes." For him, Turing did propose an actual experiment.

The most common of all charges, indeed, is that the test is committed to an operationalist or behavorist conception of intelligence, *e.g.*, Wolfe Mays (1952), Keith Gunderson (1964b), John Searle (1980) and Ned Block (1981). Mays' criticism of Turing (1952) is by far the one most loaded with anticommunist hysteria. He concluded his text with a reference to George Orwell's *1984* and "Big Brother as the Master Programmer" (p. 162). I bring this forth not

because machine-intelligence dystopias are necessarily bogus. My point here is simpler. Had he been more interested to know about Turing's background, say, as Polanyi had, perhaps his *ad hominem* attacks would have been more subtle, say, like Jefferson's (§1). Searle and Block even acknowledged that the test did not come with an operational "definition" (*i.e.*, necessary and sufficient conditions), but they then customized a notion of sufficiency behaviorism to purport that it is behavorism all the same. But Gunderson and them all produced their criticisms by means of thought experiments. They introduced into the question idealizations that look really exotic and disconnected with their opponent's frame of discussion. For comparison, I invite the reader to revisit McCarthy and Shannon's criticism of Turing's proposal, which I see as raising roughly the same point. This class of interpreters, in sum, answered "yes" to the problem. They did see the imitation game as the proposal of an actual (behavioristic) experiment. But for them it was harmful in another sense — a piece of bad science and bad philosophy, as it were.

Others answered "yes" to the dilemma but praised the test. I have reviewed views in this class elsewhere (§2.2), including the following for example. In the *causal-mental view* of Darren Abramson (2011), the test carries in it a hidden necessary condition for intelligence which he named "the epistemic-limit condition." This is, according to Abramson, the capability of the machine to cause wonder to its own designer. And in the *response-dependence view* of Diane Proudfoot (2013), the test characterizes conditions for normal observers to designate the property of intelligence under normal conditions of observation.

There have been even other interpretations, of course, whose varieties are beyond my point here. As of (2000), Saygin et al. offered a survey with broad coverage.

The idealization of the Turing test

Last but not least, there is a class of supporters of the test that see in it key interesting properties which could hardly be beaten by any alternative proposal ever suggested. Perhaps the three main representatives of this class are, in chronological order, James Moor (1976), Daniel Dennett (2006 [1984]) and Jack Copeland (2000b).

Moor's analysis was in fact a key advent in the history of the reception of Turing's test. Until then, the test had received barely any substantial defense. He offered a probabilistic view of the test (§2.2). For him, the imitation game can provide inductive evidence to eventually infer machine intelligence. With that epistemological framework in mind, Moor proposed a variant of Turing's original version of the imitation game to make it less ambiguous. In Moor's version, there is no gender question at all and there is no *man-imitates-woman* version either (we shall see Turing's original proposal soon). Here is Moor's defense of his version as an experiment design:

> [...T]he basic question which replaces the question 'Can machines think?' might be put, 'On the average after n minutes or m questions is an interrogator's probability of correctly identifying which respondent is a machine significantly greater than 50 percent?"

> If the number of minutes and questions were kept very small, then playing the imitation game would be little more than an entertaining pastime. But, in order to make the imitation game less of a game and more of a test with interesting results let us assume that the following situation occurs. The imitation game is played by many different interrogators each of whom has ample opportunity to ask many questions (each taking a week to make thousands of inquiries if you wish) and the results are such that the probability of the average interrogator correctly identifying the machine is not significantly greater than 50 percent. If such a situation did occur, then one would have little doubt that the imitation game was played well by the machine [...]. (MOOR, 1976, p. 249-50)

So Moor seems to have found in scalability the way to improve the test as an actual experiment. The question I ask at this point is this: how does Moor's scheme would look like in the eyes of a scientist? Does it seem to be anywhere near the design of an actual experiment? First, if the interrogators are to be real people, they would have to fit some of a number of admissible demographic profiles, perhaps declare biases etc.. Second, how many questions are "many questions"? Is any distribution on the number m_i of questions made by each interrogator i acceptable? How would it affect the probabilistic model? And so on.

Let us now turn to Jack Copeland and how he addressed in turn the same problem:

> It is often claimed that Turing was insufficiently specific in his description of his test. What are the specifications of a definitive test? How long? How many judges? What number of correct identifications is to be tolerated? However, these demands appear to miss the point. Whether a given machine is able to emulate the brain is not the sort of matter that can be settled conclusively by a test of brief duration. A machine emulates the brain if it plays the imitation game successfully come what may, with no field of human endeavour barred, and for any length of time commensurate with the human lifespan. Consider two time-unlimited imitation games, a man-woman game and a machine-human game, each employing the same diversity of judges that one might encounter, say, on the New York subway. If, in the long run, the machine is identified correctly no more often than is the man in the man-woman game, then the machine is emulating the brain. Any test short enough to be practicable is but a sampling of this ongoing situation. After some amount of sampling, we may become convinced that, in the long run, the machine will play as well as the man, but only because we believe that our samples of the machine's performance are representative, and we may always change our opinion on the basis of further rounds of the game. (COPELAND, 2000b, p. 530)

I think that this gives a perfect illustration of what goes on within the Turing test dilemma. I do not wish to suggest that I think that either Moor or Copeland, because they argued for the test as an actual experiment, then they absolutely must have provided a final experimental scheme for the test. My point is rather this. In spite of their best effort, there seems to be present here a sort of attitude that is not helpful in light of the dilemma. They seem to have actually believed they could anticipate from the armchair the issues and surprises one faces when designing an experiment. It turns out that Turing himself, who shifted from pure mathematics to empirical problem-solving during the war, discussed this phenomenon and remarked (1950) that there is "a fallacy to which philosophers and mathematicians are particularly subject." The fallacy, he

pondered, is to assume that "there is no virtue in the mere working out of consequences from data and general principles" (p. 451).

Dennett has been the best representative of this class of commentators. As we have seen (§3.1), he initially also defended an idealized Turing test as an actual experiment (2006 [1984]). He emphasized key conceptual properties of the imitation game as an experiment design and overlooked details that would be actually necessary to run it as a real experiment. He even defied anyone to come up with a better alternative proposal for a test for machine intelligence. However, having at first idealized the test this way, Dennett was ready to take the pains to actually engage in a committee with the goal of implementing real instantiations of the test. At the end of this experience, he admitted (2006 [1997], p. 315) to have learned about (in my words) how challenging it can be to move from conceptual experiment design to actual experiment design.

The divide over the Turing test dilemma

Thus runs the divide, *for* and *against* Turing's test to be seen and used as an actual experiment. On the one hand, the negative answer — "abandon the test, and relegate it to history (!)" — seems to throw out a core part of Turing's legacy. And Turing did make references to "[my] test." He did suggest that the test was proposed to replace "the original question" (can machines think?), which he took to be "too meaningless to deserve discussion." He did declare that "the only really satisfactory support" for answering the question "will be that provided by waiting for the end of the century and then doing the experiment." On the other hand, the positive answer — "save the test, and if needed improve it (!)" — falls short to explain, say, why Turing *did not want altogether* to "abandon the original form of the problem" (can machines think?); or, if Turing's test was really meant to be definite and practical for actual use, why did him speak of it only sketchily in 1950, and from 1951 to 1952 even make changes in its design? Now, given all the tensions between these mutually exclusive positions, one may ask: is a rapprochement possible?

I invite the reader to bear with me in the development of my argument as described before (§3.1) so that we may gain depth to revisit this question at the end (§3.8).

3.3 An interpretive basis for Turing's 1950 text

Turing's (1950) text is often said to be accessible for a general readership and yet to be complex, multi-layered and too ambiguous for scientific and philosophical interpretation if not even contradictory. In connection with the two horns of the Turing test dilemma, the core tension that has been pointed out in Turing's text can be formulated as follows. In the first part of his text Turing does seem to have meant to propose the imitation game (his test) as an actual experiment to "replace" the original question, which he suggested that was "too meaningless to deserve discussion." But then, instead of providing a precise specification of his test, he actually dedicated a central part of the text for a philosophical discussion on the original question ("can machines

think?"). How can these apparently opposing facts be reconciled? I shall start by addressing the logical structure of Turing's text as explicitly as possible in view of providing it intelligibility.

The logical structure of Turing's 1950 text

Let Turing's 1950 text be read according to this order of reasons or logical steps:

- (*The proposal*, §§§1, 2, 3). His new proposal on how best to discuss the question, "can machines can think?" One possibility, he argues, is to have the discussion on the basis of commonsense notions of machine (let us say, a steam engine) and thinking (say, what humans, and humans only do). But this, he observes, would render the question paradoxical from the start and, in effect, absurd. He poses the imitation game as an idealized scenario that he designed to be a sensible and proper substitute for the obsolete commonsense. He comments on the appeal and the settings of his proposed idealizations; in particular, why conversational question-answering by teletype makes sense as an intellectual task to empirically evaluate the capabilities of digital computers (the new machines then existing), to perform something that, if done by a human being, one ought to call "thinking." He thus proposes the imitation game as a vivid picturesque form that carried inside an (epistemological) "criterion for thinking." Based on the imitation game, he suggests two variations of the new question (the experiment) in replacement to the original one, and he will continue to suggest yet other variations of it as he proceeds into the next logical steps.

- (*The science*, §§4, 5). His teaching of what a digital computer is, in language widely accessible for readers in philosophy, mathematics and science, if not the general public altogether. He makes it clear that his proposal (above) is a philosophical reflection upon a science, namely, his 1936 mathematical science of computing. This science is now combined with the technology of stored-program computing that has been developed (early since the war years at Bletchley Park, but this he cannot reveal; and) in the postwar years. This combination was not casual, but fine-tuned by him and colleagues in order to make digital computers behave or perform as universal computing machines. His proposal, it should be clear at the end of his §5, is not about any sensible imaginary scenario but one that is informed and constrained by the science and technology of digital computers.

- (*The discussion*, §§6, 7). His negative and positive argumentation — *by means of* the science-informed proposal — against a series of nine objections to a positive answer to the original question ("can machines think?"). As preliminaries, Turing explains his beliefs and views. For him, the scientific status of the question is in the open. His own belief is that the answer for the question is positive, but would rather avoid saying it directly for the very reason why he outlined the proposal in the first place, namely, the proposal gives a basis for the discussion not to be meaningless. Then proceeding to the discussion itself he engages into each objection formulated and systematically refers to the imitation game in

> his rebuttal. This was his negative argumentation. He then advances to present a tentative research agenda for the development of "learning machines" that could be made to play the imitation game well. These, once provided with the required storage capacity, could be programmed to learn for themselves. In analogy with the education of a human child, his suggested approach was to find a "child program" that would have initially little structure but would be able to learn from experience so as to eventually exhibit intelligence of their own in the imitation game. This was his positive argumentation. Both the negative and the positive argumentation were systematically grounded on the imitation game.

Now, it is important to emphasize key elements of Turing's rationale. Without the proposal, the discussion of the original question ("can machines think?") would be grounded on commonsense notions of "machine" and "thinking" back then. From the point of view of Turing's goal of proposing a conceptual change for the meaning of these words based on a new science, this would be absurd indeed. And without introducing or teaching the new science, the proposal might be understood as some silly intellectual exercise in fiction or fantasy. But given such elemental premises, Turing reasoned, the discussion could finally unfold. It would then have a basis on his proposed (epistemological) "criterion for 'thinking,"' which was at the same time embedded in a sensible idealized scenario to keep an appeal to commonsense.

In this section I will continue to examine the internal logic of Turing's text. We shall gain more depth into it towards making the horns of the Turing test dilemma to drop out.

I will start with the second horn. Let the test be dismissed as an experiment. So this would mean that Turing's proposal boils down to a piece of rhetorics with no scientific content inside. And yet Turing does seem to have proposed the imitation game (test, or experiment) to "replace" the original question, which he suggested that was "too meaningless to deserve discussion." His reference to "meaningless" has been largely understood as affiliation to positivism, operationalism or behaviorism. I have studied Turing's attitude relative to these philosophical doctrines in depth before (§2), and clearly discouraged any connection. But then, why would Turing have referred to the original question as "meaningless"? We shall now examine it.

Turing's proposal of conceptual change on the meanings of words

In the very opening of his text, Turing made a point about the dispensability of a definition for words "machine" and "think" according to the common sense back then. He wrote:

> I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. (TURING, 1950, p. 433)

Turing thus addressed directly the issue of the paradoxical aspect of the combining words "machine" and "think." Later in the same text he even reiterated this observation in connection to the notion of "learning." He wrote: "[t]he idea of a learning machine may appear paradoxical to some readers" (1950, p. 458).

Turing's caveat, however, does not seem to have received enough attention. Wolfe Mays, for example, who was a co-participant with Turing in the 1949 Manchester seminar(s) (§A.4.2) that motivated Turing to write his 1950 paper, contributed to the discussion with one of the earliest received views on Turing's paper (1952). Mays went to the Oxford English Dictionary to promptly show that Turing had just instilled nonsense. He read from entry "machine" back then:

> [A] combination of parts moving mechanically as contrasted with a being having life, consciousness and will. Hence applied to a person who acts merely from habit or obedience to a rule, without intelligence, or to one whose actions have the undeviating precision and uniformity of a machine.
> (MAYS, 1952, p. 149, as retrieved from the O.E.D. as of 1952)

Mays seemingly had different plans and tried to push to Turing the word "robot" instead:

> The word is ready to hand and was coined by Karel Capek, we call them "robots," those devices which fall in the twilight zone between man and the normal run of machines [...]. In this connection it might be a good thing to drop the word "machine," with its emotional overtones of clanging metal, and use some such neutral word as "artifice." Machines which can perform logical and mathematical operations are very different from the steam-engines, printing-presses and looms met with in our everyday excursions. (MAYS, 1952, p. 150)

Mays' suggestion can be read from its context. Capek's "robot," coined in 1920, was a word associated with national-state tyranny through the notion of determinism. Mays, who alluded to "Frankenstein" (p. 150) and to "mechanical necromancer" (p. 153), was loaded with political, social and cultural biases. He overlooked that Turing's goal was exactly to propose to common sense a science-informed conceptual change on the meaning of words "machine" and "think." Turing had been at least since early 1947 explicitly challenging the conventional wisdom caught in common phrases such as "acting like a machine" (2004 [1947], p. 393), "purely mechanical behaviour" (2004 [1948], p. 410). The fact that even Mays, a contemporary of Turing's at the same University of Manchester, disregarded his opening plea for a suspension of judgement shows, I think very precisely, why Turing felt the need to acknowledge from the start that he understood all too well that the original question ("can machines think?") would sound absurd (or "meaningless"), and why he resorted to the imitation game as an attempt to provide a new frame of discussion hopefully towards shaking up people's intuitions about that question.

Also in 1952, right in the beginning of the BBC roundtable, University of Cambridge philosopher Richard Braithwaite acknowledged the paradoxical aspect of the original question but held in turn a very different attitude towards it:

> *Braithwaite*: We're here today to discuss whether calculating machines can be said to think in any proper sense of the word. Thinking is ordinarily regarded as so much a speciality of man, and perhaps of other higher animals, that the question may seem too absurd to be discussed. But, of course, it all depends on what is to be included in thinking. (TURING et al., 2004 [1952], p. 494)

With that, we may consider that reacting to the original question as "meaningless" or "absurd" was a standard or natural intellectual attitude back then.

So far we have gained depth into the issue by examining the intellectual environment against which Turing posed his famous phrase and reference to "meaningless" question. Now, let us look back at the internal logic of his text. In preparation to reach that point in his text, Turing pondered: "[w]e cannot altogether abandon the original form of the problem, for opinions will differ as to the appropriateness of the substitution" (1950, p. 442). He then proceeded to outline two predictions about the future, and here is where he wrote his famous phrase:

> Consider first the more accurate form of the question. 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. The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. (TURING, 1950, p. 442)

The order of exposition of Turing's sentences is striking. The next sentence, after his reference to "meaningless," refers to what he expects will happen with the combination of words "machine" and "think" in society and culture in the future. That is, it was the combination of words that Turing believed to be meaningless, not its philosophical discussion. Because such combination of words as understood according to the common sense of the time was meaningless, the question "can machines think?" was meaningless likewise. For this reason Turing thought out to replace the original question by his imitation game in the first place, but it shall be straightforward at this point that he did not mean this replacement to be literal or complete, or, to say it differently, he did want to discuss the original question. Apparently he even tried it a few times before, only to find himself being supposedly "contradicted" by the silly point that common sense did not allow any meaning to the question. The imitation game, so he seems to have reasoned, would render the original question in another form or basis that would enable its discussion. Based on the imitation game, it would be possible to articulate both the new science and common sense. Crucially, the imitation game enabled him to outline his first prediction (about a digital computer being able to play the imitation game well in the future) in differentiation with his second prediction (about the use of words and the general educated opinion in the future). Both predictions, or both points, were important for Turing, respectively, the empirical and the linguistic or pragmatic. It is precisely Turing's juxtaposition of the two predictions that shows, astoundingly, that in a certain sense, he did not want to discuss the original question, and in another sense, he dit it.

We have thus examined Turing's point about the need to replace or substantiate the original question in order to render it meaningful for discussion. Now, we understand that *Turing did want to discuss the original question in a certain sense*. But still in connection with the second horn of the dilemma, does that mean that Turing's proposal boils down to a piece of rhetorics with no scientific meat inside? Why would Turing then have referred to the imitation game as a test and experiment and considered running it as "the only really satisfactory support" (p. 455) that can be given for his views? Note that answering these questions in connection with the second horn of the dilemma requires examining *what kind of discussion Turing offered* in his 1950 text and whether or not it have scientific and philosophical significance.

In fact, Turing's discussion (§§6, 7 of his 1950 paper) turns out to have taken 19 out of 28 pages of it (nearly 70%). There should be little doubt that it was the high moment and very focus of his paper. Turing himself suggested that he proposed the imitation game as "a basis for discussion" of the original question. This appears perhaps most clearly in his discussion of the mathematical objection, when he distinguished classes of readers of his proposal:

> Those who hold to the mathematical argument would, I think, mostly be willing to accept the imitation game as a basis for discussion. Those who believe in the two previous objections [the "theological" and the "heads in the sand" objections] would probably not be interested in any criteria.
> (TURING, 1950, p. 445)

Turing's reference to "criteria" should be a clear sign that his discussion was meant to have some epistemological significance. I shall next present evidence that Turing's "discussion" was no casual initiative of his and no vacuous rhetorics. Rather he seems to have felt compelled to follow a well-known method in the history of philosophy, with a clear philosophical goal in mind.

Turing, reader of Bertrand Russell

Turing made it clear — most notably in his discussion of the theological objection (p. 443) — that he was a reader of Bertrand Russell's *History of western philosophy* (1972 [1945]), which had appeared only five years before his (1950) paper and is one of the few pieces in its bibliography. I am not aware of any commentary on the secondary literature taking notice of this, and yet, I find it to be a key exegetical element to make sense of his 1950 text. It seems that Turing turned to Russell's *History* as his chosen reference on the history of philosophy. I shall take a moment to examine this by quoting in stepwise in depth from Russell, who thus introduced the method:

> Dialectic, that is to say, the method of seeking knowledge by question and answer, was not invented by Socrates. [...] But there is every reason to suppose that Socrates practised and developed the method. [...] Certainly, if he practised dialectic in the way described in the Apology, the hostility to him is easily explained: all the humbugs in Athens would combine against him.
> (RUSSELL, 1972 [1945], p. 92)

If there was hostility to Socrates in Athens, there was hostility to Turing in postwar Britain as well (§1). Moreover, the platonic figure of Socrates, as known, was a master of irony. Accordingly the dialectic method was tailored by irony, which was one of Turing's preferred tools in debating conventional wisdom (§1.3). We may now recall Robin Gandy's anecdote about Turing reading his 1950 text to him out loud with a smile, sometimes with a giggle. It matches.

Turing did not actually mention Russell in connection with the one explicit reference that he made of "[t]he question and answer method" (1950, p. 435). In that occasion he was justifying his choice for that method as the very intellectual task to be addressed in the imitation game. So, on the hand, Turing seems to have liked Russell's notion of the method so much that he adopted it not only to debate his philosophical opponents as we shall see next, but also to test the machines in the imitation game. However, on the other hand, we will also see later that this second specific use of the method (testing the machines) is also correlated with Jefferson's citation of René Descartes, at least as much as with Russell's citation of the platonic Socrates.

Russell remarked that the socratic dialectic method, as it turns out, was *the one used by Galileo in his dialogues to advocate his theories and overcome prejudice*. And Russell pondered about the limits of the method, as exemplified with excellence by Galileo. This passage may have contributed for Turing to discover Galileo as a hero, as portrayed in the end of his rebuttal to the theological objection (1950, p. 443-4), and later in one of his 1951 BBC radio lecture "Intelligent machinery, a heretical theory" (2004 [*c.* 1951], p. 475). Russell wrote:

> The dialectic method is suitable for some questions, and unsuitable for others. [...] Some matters are obviously unsuitable for treatment in this way — empirical science, for example. It is true that Galileo used dialogues to advocate his theories, but that was only in order to overcome prejudice — the positive grounds for his discoveries could not be inserted in a dialogue without great artificiality. Socrates, in Plato's works, always pretends that he is only eliciting knowledge already possessed by the man he is questioning; on this ground, he compares himself to a midwife. (RUSSELL, 1972 [1945], p. 92-3)

The connection of this passage with Turing's argumentative approach as exhibited in the third logical part of his 1950 text (which I called "the discussion") is striking. Russell discouraged one to purport to establish any positive grounds for discoveries by means of a discussion. And along these lines, Turing wrote: "[t]he reader will have anticipated that I have no very convincing arguments of a positive nature to support my views." With a tone of irony, he completed: "[i]f I had I should not have taken such pains to point out the fallacies in contrary views." Via Russell, as it seems, Turing reproduced Socrates' approach in his negative dialectic (§6 of his 1950 text), while respecting the boundaries suggested by Galileo's in his positive dialectic (§7 of his text).

Russell resumed to consolidate his observation about the proper use of "the method of question and answer" by delimiting that it does not apply to empirical problems such as, say, "the spread of diseases by bacteria." Then he explicitly suggested the platonic-socratic method for questions about the meaning and usage of words:

> The matters that are suitable for treatment by the socratic method are those as to which we have already enough knowledge to come to a right conclusion, but have failed, through confusion of thought or lack of analysis, to make the best logical use of what we know. A question such as "what is justice?" is eminently suited for discussion in a Platonic dialogue. We all freely use the words "just" and "unjust," and, by examining the ways in which we use them, we can arrive inductively at the definition that will best suit with usage. All that is needed is knowledge of how the words in question are used. But when our inquiry is concluded, we have made only a linguistic discovery, not a discovery in ethics. (RUSSELL, 1972 [1945], p. 93)

Now, the meaning and common usage of words, namely, "machine" and "thinking," were just the central topic of Turing's 1950 paper. Nonetheless, again in line with Russell's point that no positive discovery could come out of an application of the socratic method, as we have just seen, Turing emphasized that he did not expect to have very convincing arguments of a positive nature to support his views. He rather declared that "[t]he only really satisfactory support that can be given for the view expressed at the beginning of §6" — *viz.*, his prediction about a machine being able to play well a simplified form of the imitation game — "will be that provided by waiting for the end of the century and then doing the experiment described" (p. 455). So, because Turing was in agreement with Russell's empiricist guidelines, he was compelled to acknowledge that it was only experiment that could provide satisfactory support for his views on the central question.

Also along the lines of Russell's exposition, Turing seems to have made very specific use of the dialectic method of question and answer. His goal was "to point out the fallacies in contrary views" to his hypothesis that machines can think. It is worth to highlight Russell's point about the function of this method to elicit truth after fixing "[l]ogical errors:"

> We can, however, apply the method profitably to a somewhat larger class of cases. Wherever what is being debated is logical rather than factual, discussion is a good method of eliciting truth. Suppose some one maintains, for example, that democracy is good, but persons holding certain opinions should not be allowed to vote, we may convict him of inconsistency, and prove to him that at least one of his two assertions must be more or less erroneous. Logical errors are, I think, of greater practical importance than many people believe; they enable their perpetrators to hold the comfortable opinion on every subject in turn. Any logically coherent body of doctrine is sure to be in part painful and contrary to current prejudices. The dialectic method [...] tends to promote logical consistency, and is in this way useful. But it is quite unavailing when the object is to discover new facts. (RUSSELL, 1972 [1945], p. 93)

In fact, Russell proposed the dialectic method to debate logical questions. And it was this method that Turing decided to use in his 1950 paper against his opponents on the question whether machines can think. And yet, Turing made it clear that he saw this question as an empirical rather than a logical one. He clearly acknowledged that he did not expect to settle the matter by a philosophical discussion. He rather insisted on the open status of the question from an empirical point of view. He seems also to have considered, though, that "logical errors" were the core obstacles in the way of the requisite scientific research that would lead, in the future, to the

actual demonstration of thinking machines. He addressed such logical errors in his discussion by referring to the imitation game (test, experiment) "as a basis." We shall see later (§3.7) exactly what logical errors he addressed and what scientific and philosophical significance they have.

Altogether, we have thus learned about what kind of discussion Turing offered in the third logical step of his 1950 text. In short, it was a socratic dialectic discussion that respected the empiricist boundaries indicated by Russell to have been exemplified by Galileo.

Now that we have gained more depth into facing the second horn of the Turing test dilemma, we may shift attention to the first one. Let the test be taken to be an actual experiment. Then how could we explain that it has been tried in practice as such and has been generally argued to be either underspecified or just a piece of bad experiment design? If the imitation game is seen as an experiment to "replace" the original question, what exactly is the new (experimental) question that Turing expected the imitation game to answer? We shall now look at this closely.

Turing's various imitation tests and slippery experiment design

Consider Turing's original question (Q) "can machines think?" In the beginning of his text, as known, he proposed to replace Q by another question based on an imitation game. Let us recall that Turing initially defined the game to have three roles thus distributed: "a man (A), a woman (B), and an interrogator (C) who may be of either sex" (p. 434). The interrogator's goal is "to determine which of the other two is the man and which is the woman." The goals of players A and B are in turn to deceive (A) and help (B) the interrogator in making the correct identification. On top of this initial setting, Turing formulated the first variant of the original question (Q'):

> We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?' (TURING, 1950, p. 434)

So the first variant (Q') is whether the machine's performance in a *machine-imitates-woman* game will be as good as the man's in the initial *man-imitates-woman* game. Turing specified nothing about the machine at this point. After Judith Genova (1994), Patrick Hayes and Kenneth Ford (1995) called this version of the imitation game "the gender test." The interrogator's task would be solely to identify who is the woman, without being aware or having the bias that one of the initially unidentified entities could be a machine rather than a man. So, to pass this version of the game, the machine has to be able to impersonate a woman, that is, it has to be able to imitate gender. Supporters of Turing's 1950 proposal such as Gualtiero Piccinini (2000) and James Moor (2001) have been unable to find scientific or philosophical significance in this version of the game associated with variant Q'. They then sought to dismiss it in favor of the third variant of the question (Q''') that Turing introduced further on in his text, as we will see next. In general, this class of interpreters did acknowledge that Turing himself proposed several variations of his

imitation game, but they proceeded as if there had to be one best variant of the imitation game outlined by Turing that should be chosen over the others. Their struggle with Turing's own text is intelligible if they had the goal of finding in it one specific experiment design to hold on for.

After posing the first variant Q', there are exactly two other explicit pauses or transition moves that Turing made in the logical development of his argument (at the endings of §3 and §5). He thereby presented two additional variants of the original question. Observe that these two moments separate the three logical steps or parts that I pointed out above in my proposed reading of Turing's text (§3.3), namely, at the endings of the "proposal" part and the "science" part.

At the end of his section §3, Turing paused and made this critical observation:

> There are already a number of digital computers in working order, and it may be asked, 'Why not try the experiment straight away? It would be easy to satisfy the conditions of the game. A number of interrogators could be used, and statistics compiled to show how often the right identification was given.' (TURING, 1950, p. 436)

He then asked a second variant (Q''): "we are not asking whether all digital computers would do well in the game nor whether the computers at present available would do well, but whether there are imaginable [digital] computers which would do well" in the imitation game (p. 436), and it is here where both the concept of a "digital computer" and the (implicit) hypothesis on the existence of a certain machine (digital computer) in the future firstly appear. Turing would then dedicate "the science" part of his text just to teach what a digital computer is. This was an important move that he only could make after having introduced "the machines concerned in the game" (p. 435), which completed "the proposal" part of the text.

The second pause and transition was at the end of his section §5, when he asked a third variant (Q'''), whether there can be a digital computer C that, once enabled with "adequate storage," "speed of action" and "an appropriate program", can "be made to play satisfactorily" the "part of A in the imitation game, the part of B being taken by a man?" (p. 442). So this one makes up a *computer-imitates-man* version of the game. After Genova (1994), Hayes and Ford (1995) called this version of Turing's imitation game "the species test." It is important to observe the conditional. It is not only about the technology, but also about the science. The machine will need to have "an appropriate program." (I will study what Turing meant by that in detail in §2.)

Subsequently still in the same moment of his text which lasts until the formulation of the theological objection, Turing asked yet another question (Q''''). He asked whether "in about fifty years' time it will be possible to programme computers" with a certain storage capacity to play the imitation game to achieve a certain performance (deceive the interrogator in at least 30% of the times) when the game is played for a short time (halted after five minutes of questioning). Turing stated his belief that this fourth variant Q'''' will be decided positively. So he chose this simplified form of the game as a basis for making an empirically decidable prediction.

It is the third variant (Q'''), however, that supporters of the test appreciate the most. They construe a *computer-imitates-human* version of the game, sometimes combining it with the first variant of the question in order to use the *man-imitates-woman* version of the game as a baseline to compare results against. They argue that in the passage when Turing introduced the third variant of the question he referred to "man" in the species sense, meaning a human being in general. This is in spite of the fact that Turing had made it quite clear from the beginning that he was thinking about gender. So why would he all of a sudden shift to a species sense of "man" without further notice? The cornerstone of the argument for this interpretation, as I understand it, evokes the view that overall in Turing sources it is clear that his goal is to propose a test for machine intelligence by taking the human species intelligence as a reference. I think that this is true as an overall interpretation of Turing sources, and perhaps even obvious. What is not obvious at all is an implicit assumption of these interpreters. They take for granted that there must be some best version of the imitation game to be selected to figure as an actual experiment, in spite of the glaring fact that Turing presented several variants of it apparently with no commitment at all neither to mathematical nor to experimental precision in the design of any one of them. In fact, Turing articulated various versions of his imitation game *before, during, and after the development of his 1950 paper*. At the end of his 1950 paper, in the same way as he had discussed in (2004 [1948], p. 420-1), Turing was not even sure about which intellectual field was best to explore and test for machine intelligence. He thus wrote:

> We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. 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. [...] Again I do not know what the right answer is, but I think both approaches should be tried.
> (TURING, 1950, p. 460)

And in fact, up to the last thoughts of his on machine intelligence that we have on record, in *c.* late 1952 (§A.4.8), Turing kept both chess-playing and the unrestricted (topic-free) viva-voce examination as interesting intellectual tasks to test for machine intelligence.

Before that, in the January 1952 BBC roundtable Turing had given yet another description of "[his] test" (p. 495). As in 1950, it was based on a unrestricted viva-voce examination but then Turing invoked the scenario of a law court with the machine being interrogated by a jury:

> You might call it a test to see whether the machine thinks, but it would be better to avoid begging the question, and say that the machines that pass are (let's say) 'Grade A' machines. The idea of the test is that the machine has to try and pretend to be a man, by answering questions put to it, and it will only pass if the pretence is reasonably convincing. A considerable proportion of a jury, who should not be expert about machines, must be taken in by the pretence. They aren't allowed to see the machine itself — that would make it too easy. So the machine is kept in a far away room and the jury are allowed to ask it questions, which are transmitted through to it: it sends back a typewritten answer.

> [...] Well, that's my test. Of course 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*. It's not the same as 'Do machines think,' but it seems near enough for our present purpose, and raises much the same difficulties. (TURING et al., 2004 [1952], p. 495, emphasis added)

Turing explicitly stated that an answer to his slippery experimental question does not have to be decisive with respect to the original question on whether machines can think. He also explicitly posited what his suggestion is really about, namely, the new question, either in this particular version or, as I interpret, in any one of the several other versions he had presented before, should serve as *a basis for the discussion of the original question*. Regarding the time when a machine would be able to perform well in the jury-based version of his test, Turing conceded to Newman that it would take "at least 100 years" from the early 1950's. (Sometimes this is referred in the literature as if Turing had postponed the target date of his 1950 prediction. However, the versions of the imitation game or test at stake in each prediction are *not* the same.)

Now, I shall conclude this examination of Turing's multiple versions of his imitation game or test by bringing forth yet another key passage. To my knowledge it has not yet been noticed in the secondary literature. Still in the BBC roundtable, at some point Turing said:

> This means that if the machine was being put through *one of my imitation tests*, it would have to do quite a bit of acting, but if one was comparing it with a man in a less strict sort of way the resemblance might be quite impressive. (TURING et al., 2004 [1952], p. 503, emphasis added)

So Turing himself explicitly acknowledged that he had been (from 1948 to 1952) presenting and discussing several "imitation tests." Altogether, I conclude, the view that Turing had proposed some specific, well-defined test for machine intelligence turns out to be far fetched.

We have so far studied in some depth in what sense Turing proposed to "replace" his original question ("can machines think?"). Now, it shall also be in order for a better understanding of Turing's proposal that we analyze the basic epistemological structure of this question. We will also pay more attention, from now on, to the fact that his proposal came out of a public controversy that was started in a newspaper with Geoffrey Jefferson and was further developed at the University of Manchester in 1949 (§§A.4.1, A.4.2).

3.4 "Machines can think" implies an existential hypothesis

The positive proposition "machines can think" implies a related proposition that has the logical structure "there exists x," where x stands for a thinking machine. This logical proposition, for Turing as we have seen (§3.3), carried some empirical content inside. Its truth value thus was supposed to be decided empirically. To say it differently, whoever states the proposition, either in its positive or negative form, makes an empirical knowledge claim. It is in this sense that we shall study in this section not the logical, but the epistemological structure of that proposition.

The "machines" Turing considered were the newly conceived *digital computers* (1950, p. 436). They can be "classified," Turing added, "amongst the 'discrete state machines"' (p. 439). Turing also informed that digital computers could be considered to actually exist as of 1950:

> The reader must accept it as a fact that digital computers can be constructed, and indeed have been constructed, according to the principles we have described, and that they can in fact mimic the actions of a human computer very closely. (TURING, 1950, p. 438)

Let us now make a step back. Earlier in (1936), Turing had defined the abstract machines that we call today, after him, Turing machines. As a special kind of such machines, Turing also defined the so-called universal computing machines. (For an introduction, the reader may refer to §A.2.1.) He did so at least twelve years before the construction of the Manchester "Baby" computer in 1948, which can be considered the first instance of a universal machine to be known to exist in the real world (§A.3.8). For Turing (and for Max Newman too, as pointed out in 1949), the mathematical existence of a thinking machine could be sort of a corollary to the mathematical existence of a universal machine as shown by Turing in 1936. If the "Baby" was an empirical realization of the universal machine, then the actual existence of a thinking machine would depend only on providing it with enough storage capacity and a suitable instruction table or program to make it do something that can be called thinking. Tasks such as code breaking, chess-playing, the learning of languages and their translation, the solving of mathematical problems and so on could all in principle be performed by a universal machine by insertion of the proper special-purpose instruction table or program. Nonetheless, which of such tasks could really qualify as thinking? Could they be encoded into a program? These were precisely the points of contention in the start of the Jefferson-Turing controversy (§A.4.1). Jefferson denied that a machine could think. Newman presented a note in view of streamlining part of the confusion (1949). As we have seen in the Introduction, he formulated a distinction between the universal predication ("Can *all* kinds of thought, logical, poetical, reflective, be imitated by machines?") and the existential predication ("Can *anything* that can be called 'thought' be so imitated and, if so, how much?", no emphasis added) of tasks that could qualify as thinking. I will study Turing's hypothesis on the existence of a thinking machine in more depth later (§2). In any case, fom the point of view of the epistemological structure of Turing's existential hypothesis, it shall be clear that it is Newman's existential predication that matters. For Turing, with no loss in generality, a machine thinks if it can imitate *something* that can be called "thought."

We may go a bit further in this digression and distinguish the existence of a machine (a *bona fide* kind and sortal entity) from a property of it such as to have intelligence or thinking. Turing's hypothesis might then find better suit if formulated with additional logical structure, say, "there exists x such that x has property p," where x stands for a machine (digital computer) and p for the property of thinking or being able to think. One might try to argue that some specific question of the type "does x exist?" (depending on how x is defined, if defined at all) is not well

posed or does not make sense, and this was also at stake in the controversy on the question of the existence of a thinking machine. None of that, however, is key detail to my purpose in this section. I am concerned with the general epistemological structure of Turing and Jefferson's positions with respect to the original question "can machines think?" and how their related propositions can be evaluated empirically. It will simplify matters then, with no loss in generality, to consider Turing's hypothesis as a proposition of type "there is or can be x," where x stands for a thinking machine. Jefferson in turn, who was actually the first of the two to go publicly, was the proponent of its negation "there can be no such thing as x." We may now proceed.

A constraint on the empirical evaluation of existential hypotheses

It is an epistemological fact that *existential hypotheses cannot be refuted* — no one run or set of runs of the imitation game or whatever experiment one may think of will ever be able to prove that it is *not* the case that machines can think. The reason is that hypotheses of this kind have an open-ended nature. To show that no x can ever exist in nature would require an impossible task, namely, to search the whole space-time in order to establish that something does not exist, has never existed, and will never exist. The only definitive answer that can be given to such a question is a positive one. That can be done by presenting an instance of x (or evidence of its presence) that satisfies some agreeable notion of what such an instance or evidence should be. Otherwise, the question will remain always open for further inquiry. On the other hand, accordingly, *the negation of an existential hypothesis cannot be confirmed* — by the same logic, to show that "there is or can be no such thing as x" one has to establish that it does not exist, has never existed, and will never exist. Karl Popper famously articulated those points in his *Logic of scientific discovery* (2002 [1959]), in the context of his analysis of "strictly universal statements" and "strictly existential statements" (p. 47-8). Popper, however, went much further. He was concerned with posing a criticism of logical positivism and proposed a demarcation criterion (falsificationism) in which the empirical sciences should only be concerned with universal laws (prohibitions), as these can take the form of "strictly universal statements" and then can be refuted or, as he called, falsified. So the confirmation of existential hypotheses, for Popper, is not really part of the scientific enterprise unless the hypotheses can be transformed to some universal statement to be made falsifiable. I shall introduce and discuss in depth later (§2) a broader view due to Herbert Feigl on the interpretation of existential hypotheses in the empirical sciences, which is also critic of logical positivism (1950). For the purpose of my specific and simpler analysis of the epistemological structure of existential hypotheses in this chapter, neither Popper nor Feigl's views are actually needed. Turing himself held related insights, as I show next.

The epistemological asymmetry above can also be expressed in connection with the concept of scientific induction. For instance, suppose one is shown a first machine, a second, a third and so on, and none of them can be said to think. One may then be tempted to conclude by induction that no thinking machine exists or can exist. But this conclusion, as known, is not

warranted, and this leads us back to the fact that the negation of an existential hypothesis can never be confirmed. Accordingly, it turns out that Turing himself had an understanding of this epistemological asymmetry. He firstly commented on it explicitly in his 1948 formulation of objection (c) which became in 1950 his objection (5), the argument from various disabilities:

> A man has seen thousands of machines in his lifetime. From what he sees of them he draws a number of general conclusions. [...] A few years ago, when very little had been heard of digital computers, it was possible to elicit much incredulity concerning them, if one mentioned their properties without describing their construction. That was presumably due to [an...] application of the principle of scientific induction. These applications of the principle are of course largely unconscious. When a burnt child fears the fire and shows that he fears it by avoiding it, I should say that he was applying scientific induction. (I could of course also describe his behaviour in many other ways.) The works and customs of mankind do not seem to be very suitable material to which to apply scientific induction. A very large part of space-time must be investigated, if reliable results are to be obtained. Otherwise we may (as most English children do) decide that everybody speaks English, and that it is silly to learn French. (TURING, 1950, p. 448)

Turing returned to it when rebutting objection (8), the argument from informality of behavior:

> [...W]e cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct. The only way we know of for finding such laws is scientific observation, and we certainly know of no circumstances under which we could say, 'We have searched enough. There are no such laws.' (TURING, 1950, p. 452)

I take from these passages that Turing was at least in part aware about the epistemological asymmetry that constrains the empirical evaluation (and decidability) of existential hypotheses.

In short, overall, Turing's position can be confirmed but not refuted. Jefferson's position in turn cannot be confirmed but only refuted. In fact Jefferson's position cannot be empirically established, only defeated. He could do nothing else but propose some empirical constraint such that, if not satisfied, his position would remain defensible from a philosophical point of view. Turing's position could in principle be established empirically as long as he committed to an agreeable notion of what a positive instance of thinking machine should be. So at this point we may ask: what did Turing actually do? This leads us to recapitulate the state of Turing's scientific research on machine intelligence as of 1948, before his controversy with Jefferson was started.

Turing's choice of intellectual task to test machine intelligence as of 1948

As implied in Newman's distinction, the intellectual task addressed by a machine in a test will be critical for an inter-subjective attribution of intelligence to it. The choice of this task to compose some experiment design is thus key to open the way for an empirical confirmation of machine intelligence as conjectured in Turing's hypothesis. It turns out that early from Turing's wartime service in 1941, and most clearly in late 1945 — when Turing wrote that "chess requires

intelligence" — on up to at least the summer of 1948, Turing worked with the game of chess as intellectual task to illustrate, develop and test his concept of machine intelligence (§A.3).

In fact, in the summer of 1948 as we note from his report to the National Physical Laboratory (§A.3.9), Turing was still unable to use the Manchester "Baby" machine for his experiments. He thus created in collaboration with his friend, statistician David Champernowne, the notion of a "paper machine" (2004 [1948], p. 416). This was a scheme designed for a human being to simulate a machine in playing chess. Turing wrote back then: "[p]laying against such a machine gives a definite feeling that one is pitting one's wits against something alive" (p. 412). Most importantly here, in any case, is that Turing described the design for an initial experiment on machine intelligence, as yet unnamed, that he said he had actually done at that point:

> It is possible to do a little experiment on these lines, even at the present stage of knowledge. It is not difficult to devise a paper machine which will play a not very bad game of chess. Now get three men as subjects for the experiment A, B, C. A and C are to be rather poor chess players, B is the operator who works the paper machine. (In order that he should be able to work it fairly fast it is advisable that he be both mathematician and chess player.) Two rooms are used with some arrangement for communicating moves, and a game is played between C and either A or the paper machine. C may find it quite difficult to tell which he is playing. (This is a rather idealized form of an experiment I have actually done.) (TURING, 2004 [1948], p. 431)

The first and the last phrases are significant. The game of chess had made it "possible" for Turing to actually do an experiment on machine intelligence. The passage also shows that Turing was conscious about setting "idealized" conditions for the design of an experiment in the sense of Ernst Mach as we have seen (§3.3). Finally, note also that important elements of the structure of the (yet unnamed) imitation game were present at that point in the summer of 1948: (i) three players assume different roles in the game; (ii) there is a criterion to measure how successful the machines was; and (iii) the game allows for one to have (or not) a feeling to be interacting with "something alive." These elements will be preserved in Turing's 1950 imitation tests.

We also know that Turing has given a lot of thought about which intellectual task to choose for. In the same 1948 source, he listed five "branches of thought" or "fields" he had been considering as domain sources to pick some intellectual task from:

> [...In] setting about our task of building a 'thinking machine' [...we] are then faced with the problem of finding suitable branches of thought for the machine to exercise its powers in. The following fields appear to me to have advantages:
>
> (i) Various games e.g. chess, noughts and crosses, bridge, poker
> (ii) The learning of languages
> (iii) Translation of languages
> (iv) Cryptography
> (v) Mathematics. (TURING, 2004 [1948], p. 420)

Considering the five fields, Turing chose for the field of games (and chess specifically) over the others fields (ii, iii, iv, v). He presented his rationale:

> Of these (i), (iv), and to a lesser extent (iii) and (v) are good in that they require little contact with the outside world. For instance in order that the machine should be able to play chess its only organs need be 'eyes' capable of distinguishing the various positions on a specially made board, and means for announcing its own moves. Mathematics should preferably be restricted to branches where diagrams are not much used. Of the above possible fields the learning of languages would be the most impressive, since it is the most human of these activities. This field seems however to depend rather too much on sense organs and locomotion to be feasible. (TURING, 2004 [1948], p. 421)

It is remarkable to see that, as of the summer of 1948, Turing had considered among his choices field (ii), the "learning of languages," which we can associate with the task of conversational question-answering as addressed in his 1950 paper. He was aware about the greater appeal of this field to showcase machine intelligence, as it is "the most impressive, since it is the most human of these activities." He happens to have offered, however, an argument based on convenience and feasibility, which can be related to two main factors. First, Turing wanted to avoid technicalities required to simulate the human perception and locomotion system (including the cost of acquiring or developing artificial sensory-motor organs), about which he wrote: "[t]his would of course be a tremendous undertaking" (2004 [1948], p. 420). Second, Turing understood that the more restricted the field of an intellectual task is, the lesser is the demand for memory or the storage capacity of the machine. In 1947, he had said:

> [...] The memory capacity of the human brain is probably of the order of ten thousand million [10^{10}] binary digits. But most of this is probably used in remembering visual impressions, and other comparatively wasteful ways. One might reasonably hope to be able to make some real progress with a few million [10^6] digits, especially if one confined one's investigations to some rather limited field such as the game of chess. It would probably be quite easy to find instruction tables which would enable the ACE to win against an average player. Indeed Shannon of Bell Telephone laboratories tells me that he has won games playing by rule of thumb. But I would not consider such a victory very significant. What we want is a machine that can learn from experience. (TURING, 2004 [1947], p. 393)

Altogether, it is clear that Turing thought carefully about the main issues involved in the selection of a specific intellectual task for exploring and testing for machine intelligence. Up to the late 1940's, for reasons of convenience and short-term feasibility, chess was his preferred choice. Turing had noticed, as of 1948, about the greater appeal of "the learning of languages." But he pondered that an intellectual task in this field would be too costly to engineer, so much so that it would not be "feasible." For Turing, therefore, the choice for a particular intellectual task to function as a standard for thinking was a matter of trading the appeal or potential of the task to cause wonder (§2.3), on the one hand, by convenience and feasibility, on the other hand. (For short, henceforth, the appeal *v.* feasibility tradeoff.) Moreover, by the very fact that various intellectual tasks were considered by Turing as standards for a machine intelligence test, we can even establish that, conceptually, his imitation tests can be task-independent or task-free.

Turing could have sticked to his cost-benefits rationale and been, like Claude Shannon (cited above), one more computer pioneer to have referred only to chess and other narrow tasks. Indeed, as he himself argued, chess had several advantages. And yet, in his 1950 paper Turing opted to address conversational question-answering as intellectual task, which was clearly far beyond the possibilities of the time. Now, this choice killed any chance for Turing to empirically test his hypothesis in his lifetime. And he knew it very well, as evidenced by his predictions with target dates to at least fifty years from his time. So at some point he changed his mind and made a significant move to choose instead for field (ii), the learning of languages, which he saw as the most appealing yet an infeasible one back then. Why did Turing make that move? The reason, as historical evidence suggests, lies in the controversy with Jefferson and others over the question "can machines think?" We shall now look into it in order to gain insight into the conceptual problems that Turing felt himself compelled to address in his 1950 imitation tests.

3.5 1949, the crucial year

Let us briefly recapitulate Turing's moves ever since his experimental turn during the Second World War (§A.2.2). In 1945, just after the war, he had joined the National Physical Laboratory (NPL for short) to "build a brain" (§A.3.2). To do that, of course, he depended on collaboration with electrical engineers. But he was unhappy with the slow progress on the NPL computer project relative to rival projects in the US and in Britain, and seems to have lost trust in the organizational and intellectual environment at that institution (§A.3.7). In the summer of 1948, he left the NPL to join Newman's Computing Laboratory at the University of Manchester and lead the design of software to be run on the first and only existing implementation back then of a universal computing machine (§A.3.8). Furthermore, early from Turing's wartime service in 1941 and most clearly in late 1945 on up to at least the summer of 1948 (§A.3), Turing had chosen and was actually working with the game of chess. This was for him, as we have just seen (§3.4), the best feasible intellectual task to illustrate, develop and test machine intelligence. My point here, in sum, is that what we see in a chronology of Turing's moves is a pursuit of the best conditions for *physical experiment*. It seems to have come in the way of it from 1949 on, though, that his access to the Manchester "Baby" computer turned out not to be as easy as he may have imagined. As Newman is reported to have said, "the engineers took over" the Manchester computer (2006, p. 187; §A.3.8). So the prospect for Turing to run experiments on a real computer was not promising. A crucial factor to make Turing shift from physical to *thought experiment*, as I have been suggesting nonetheless, is that he engaged in a public controversy that sparked deeper philosophical discussions. Outraged by strong claims, he was led to react.

It all begun in June 1949. Computer pioneer Douglas Hartree was publishing his *Calculating instruments and machines*, in which the new electronic computing machines could do a lot and yet should be seen as nothing but calculation engines. Distinguished neurosurgeon Geoffrey Jefferson had given his Lister Oration along the same lines, and pushed strong demands to accept

that "machine equals brain." Asked by a reporter from *The Times*, Turing rebutted to Jefferson sharply, in wit. This indirect exchange with Jefferson, however, as it seems, would only make an impact on Turing's views after the 27th of October 1949. On this date they both participated in a seminar on "Mind and computing machine" in the Department of Philosophy of their university co-chaired by Michael Polanyi, who also became an (even if but the gentlest) intellectual opponent of Turing's. These three conservative thinkers, all endowed with fellowships of the Royal Society and prestigious university professorships, tried to establish boundaries to Turing's views on machine intelligence. From June to December 1949 their provocations would resonate in Turing's thought and crucially lead to his famous 1950 paper. I shall now reconstruct it.

Turing provoked by Douglas Hartree

Douglas Hartree (1897-1958), Fellow of the Royal Society since 1932 (DARWIN, 1958), then Plummer Professor of Mathematical Physics and member of the Cavendish Laboratory at the University of Cambridge had given his "short series of lectures" in the early fall of 1948 at the University of Illinois. His related *Calculating instruments and machines* came out in about June (1949). Hartree had cited in his May 1949 preface (p. v) the Manchester Baby computer, which had recently been "put into operation." (Earlier, in February 1946 Hartree had been a key figure for Newman's Computing Laboratory in Manchester to be granted funding from the Royal Society, cf. §A.3.8.) And he kept pushing his public criticism on the term "electronic brain" (1949, p. 70) as he had been doing ever since his note on *The Times* in early November of 1946 in the wake of Louis Montbatten's address (§A.3.3). It is after Hartree that Turing cited and discussed "Lady Lovelace's objection" (1950, p. 450) or "dictum" (2004 [1951], p. 485). (The reader may observe that, in his 1948 report, Turing had discussed the same view or objection "(e)," cf. my Fig. 1, with no mention of Lady Lovelace at all.) Hartree drew attention to her views:

> Some of her [Lady Lovelace's] comments sound remarkably modern. One is very appropriate to a discussion there was in England which arose from a tendency, even in the more responsible press, to use the term "electronic brain" for equipment such as electronic calculating machines, automatic pilots for aircraft, etc. I considered it necessary to protest against this usage [Hartree, D. R. *The Times* (London), Nov. 7, 1946.], as the term would suggest to the layman that equipment of this kind could "think for itself," whereas this is just what it cannot do; all the thinking has to be done beforehand by the designer and by the operator who provides the operating instructions for the particular problem; all the machine can do is to follow these instructions exactly, and this is true even though they involve the faculty of "judgment." I found afterwards that over a hundred years ago Lady Lovelace had put the point firmly and concisely (C, p. 44): "The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever *we know how to order it* to perform" (her italics). (HARTREE, 1949, p. 70)

Hartree resumed it with this passage, which conceded a window for machine learning research:

> This does not imply that it may not be possible to construct electronic equipment which will "think for itself," or in which, in biological terms, one could set up a

> conditioned reflex, which would serve as a basis for "learning." Whether this is possible in principle or not is a stimulating and exciting question suggested by some of these recent developments [...]. But it did not seem that the machines constructed or projected at the time had this property. (HARTREE, 1949, p. 70)

This passage would be quoted and discussed by Turing at length (1950, pp. 450, 454, 459). Turing was decided to pursue machine learning beyond "reflexes" and "the action of the lower centres" of the brain at least since his *c.* November 1946 letter to Ross Ashby (§A.3.4). In fact, as we will now see, in 1949 Hartree was already writing in reply to Turing.

When the Jefferson-Turing controversy broke out on *The Times* 11 June 1949, Hartree jumped into it with letters to *The Times* that appeared in "The Mechanical Brain" correspondence on 11 June 1949 (p. 4) and on 16 June 1949 (p. 2). Back in November 1946, he had been interviewed alongside Turing about the machine (or "brain") under construction at the NPL, the so-called Automatic Computing Engine (ACE). The *Daily Telegraph* reported on 7 November "'ACE' will speed jet flying," an account based on their interviews but under headline centered on Hartree's views. He would have said: "[t]he implications of the machine are so vast that we cannot conceive how they will affect our civilisation." But he meant practical applications of scientific computing. Turing would have gone his own way: "Dr Turing, who conceived the idea of [ACE], said that he foresaw the time, possibly in 30 years, when it would be as easy to ask the machine a question as to ask a man." The contrast between Hartree's view and Turing's view was marked. Hartree has also been reported to have said in that 1946 interview, in line with his future citations of Lady Lovelace, that "the machine would always require a great deal of thought on the part of the operator." And he would have denied "any notion that Ace could ever be a complete substitute for the human brain," perhaps also because he feared that "[t]he fashion which sprung up in the last 20 years to decry human reason is a path which leads straight to Nazism."

Apart from Hartree's play of the Nazi card (§1.6), Turing must have felt himself compelled to respond to what I shall also call henceforth the *Lovelace-Hartree thesis*. Soon after their November 1946 interviews, in his February 1947 lecture to the London Mathematical Society Turing had already defined what would be his line of response. He accepted a premise of the thesis, and questioned its conclusion:

> It has been said that computing machines can only carry out the processes that they are instructed to do. This is certainly true in the sense that if they do something other than what they were instructed then they have just made some mistake. It is also true that the intention in constructing these machines in the first instance is to treat them as slaves, giving them only jobs which have been thought out in detail, jobs such that the user of the machine fully understands what in principle is going on all the time. Up till the present machines have only been used in this way. But is it necessary that they should always be used in such a manner? (TURING, 2004 [1947], p. 392-3)

Turing seems to have observed that the objection raised by the Lovelace-Hartree thesis was strong and could only be met if machines were made to *learn* for themselves by experience, with

no need to redesign. He said: "[w]hat we want is a machine that can learn from experience." And completed: "[t]he possibility of letting the machine alter its own instructions provides the mechanism for this" (2004 [1947], p. 393). So, when Hartree wrote the above passage in (1949, p. 70) denying that "the machines constructed or projected at the time had this property" (of learning to think for themselves), he was already responding to Turing (February 1947), and maybe also to Norbert Wiener (October 1948) after Turing. (For a brief review of the Wiener-Turing connection, the reader may refer to §A.3.6.) For in his famous *Cybernetics* (published on 22 October 1948), Wiener made a specific citation of Turing's results from his 1936 paper (1965 [1948], p. 125-6) to conclude that "the logic of the machine resembles human logic, and, following Turing, we may employ it to throw light on human logic." From that passage, Wiener proceeded to answer positively to the possibility of the machine to have even "a more eminently human characteristic," namely, "the ability to learn." Wiener made it public thereby that he shared Turing's non-obvious view that machines could be made to learn for themselves. In fact, as we will see soon in this section, Wiener's *Cybernetics* did not pass unnoticed in Britain and may have contributed to provoke Hartree's reaction.

It turns out, as we shall see in a bit more detail later (§3.7), that the imitation game embodies Turing's response to the Lovelace-Hartree thesis in its very design. Turing observed that for a machine-intelligence experiment to be effective against that thesis, it would need to showcase that machines can learn from experience like we humans do. Now, conversational question-answering provides a general and convenient illustration scenario for just that, for if machines can learn directly from experience then they would have to be able to respond immediately to new information submitted to it by an interrogator.

Turing provoked by Michael Polanyi

Michael Polanyi (1913-1976), born Hungarian, left Nazi Germany to England to become Fellow of the Royal Society in 1944 (WIGNER; HODGKIN, 1977). In 1948, associated with the Department of Philosophy and with some support from Professor of Philosophy Dorothy Emmet, he was granted a new chair of Social Studies at the University of Manchester. (Emmet, as an Alfred Whitehead scholar, was interested in science and had religious affinities with Polanyi. For details, see Jonathan Swinton's 2019 study, p. 87-90). Emmet and Polanyi were interested in the postwar public discussion about science and society, and payed attention to the debate around the new electronic machines or "electronic brains" (§§A.3.3, A.4.1). So they invited Turing, Newman, Jefferson and others to a seminar on "the mind and the computing machine," to be held on 27 October 1949 at the philosophy department. This would be a crucial event. We know of it mostly from minute notes that survived (TURING et al., 2005 [1949]). According to Wolfe Mays, they were taken by some unidentified member of the department, who Swinton guessed to have been Desmond P. Henry (2019, p. 92). The seminar had two sessions. Here I will cover some of Polanyi's interventions. (For more on the two sessions, the reader may refer to §A.4.2.)

The first session was led by Polanyi, who read a text, entitled "Can the mind be represented by a machine? Notes for discussion on 27th October 1949," which he had prepared and circulated to Newman and Turing several weeks before the meeting. Essentially, Polanyi claimed that humans can solve problems that machines cannot. He vindicated support from Gödel's incompleteness theorems. (For an overview of them, see §A.2.1). Polanyi scholar Paul Blum examined a printed copy of that text which is available at the Polanyi archive at the University of Chicago. Blum noted annotations that may have been made by Turing. He wrote (2010, p. 52) that there were a few corrections that were "certainly by Polanyi," but there were "three comments by a different hand," and speculated: "compared with manuscripts published at the Turing Digital Archive (www.turingarchive.org) they could be by Turing." I reproduce below what those three comments (possibly by Turing) are:

> The discoveries of Gödel (1930) have shown that arithmetic and advanced geometry are incomplete. [Ed. Superscript by unknown hand: *rather: number theory*.] (BLUM, 2010, p. 52, note 40)
>
> There is established thus an inexhaustible procedure for the discovery of ever more true mathematical formulae, which, by its very nature, is incapable of formalisation. [Ed. "nature... formalization": underlined and annotated by unknown hand: *no./not in the same language. But we can formalize the meta-language.*] (BLUM, 2010, p. 52-3, note 41)
>
> Our minds however are not similarly limited. [Ed. Superscript by unknown hand: *But they are. Otherwise we get into the paradoxes*.]
> (BLUM, 2010, p. 53, note 44)

These annotations are consistent with what Turing is reported to have said in the seminar itself and also with statements of his elsewhere. In the first minute notes of the seminar, we read:

> NEWMAN TO POLANYI: The Gödel extra-system instances are produced according to a definite rule, and so can be produced by a machine. The mind/machine problem cannot be solved logically; it must rest on a belief that a machine cannot do anything radically new, to be worked on experimentally. The interesting thing to ask is whether a machine could produce the original Gödel paper, which seems to require an original set of syntheses.
>
> TURING: emphasises the importance of the universal machine, capable of turning itself into any other machine.
>
> POLANYI: emphasises the Semantic Function, as outside the formalisable system. (TURING et al., 2005 [1949])

This gives evidence that Newman considered, just like Turing as we will see in depth later (§2), that "the mind/machine problem" can be decided empirically and only empirically. That is, for Newman as well, it is not merely a language problem as is sometimes suggested (cf. §3.2). More than that, Newman shifted the discussion around Polanyi's Gödelian argument to the Lovelace-Hartree thesis. So, apart from correcting mistakes in Polanyi's interpretation of Gödel, Turing and Newman seem to have tried to extract some philosophical meat from Polanyi's point. Specifically, Newman had cast the problem of "produc[ing] the original Gödel paper" as an instance of Lady

Lovelace's objection (the question whether a machine can "do anything radically new"). And this had been suggested by Turing himself ever since his February (2004 [1947]) lecture, when he connected his response to (then yet unnamed) Lady Lovelace's objection (p. 392-3) — machine learning — with his response to Gödel's argument or the mathematical objection (p. 393-4). It is worth noting that the problem of "produc[ing] the original Gödel paper" would reappear in the secondary literature, pushed first by John Lucas (1961) and later by Roger Penrose (1989), and drive a lot of discussion. To my knowledge, it was only in 2005 that Turing scholars had access to the survived minute notes of the October 1949 Manchester seminar (§A.4.2). Lack of historical knowledge seems then to have prevented Lucas and Penrose to have observed that Turing (and Newman), who were key character(s) for their discussion, had themselves already responded to what is perhaps the core aspect of their uses of the so-called "Gödelian argument."

Polanyi's appeal to a "Semantic Function" would be extended into the second session of the seminar, chaired by Dorothy Emmet, and lead to new exchanges with Turing. At some point, we see in the notes that Turing would have presented a distinction, to which Polanyi replied:

> TURING: declares he will try to get back to the point: he was thinking of the kind of machine which takes problems as objectives, and the rules by which it deals with the problems are different from the objective. Cf. Polanyi's distinction between mechanically following rules about which you know nothing, and rules about which you know.
>
> POLANYI: tries to identify rules of the logical system with the rules which determine our own behaviour, and these are quite different things.
> (TURING et al., 2005 [1949])

Here lies the motivation for Turing's (1950) formulation and rebuttal of the "argument from informality of behaviour" (p. 452), which shows that Turing felt compelled to respond to Polanyi. In fact, as we shall see (§3.7), the imitation game embodies Turing's response in its very design.

Writing nine years after the seminar, Polanyi gave this valuable piece of historical information about the first session:

> A. M. Turing has shown [Polanyi's note: in a communication to a Symposium held on "Mind and Machine" at Manchester University in October, 1949. This is foreshadowed in 'Systems of Logic based on Ordinals', *Proc. London Maths. Soc.*, Series 2, 45, 1938-9, pp. 161-228.] that it is possible to devise a machine which will both construct and assert as new axioms an indefinite sequence of Gödelian sentences. Any heuristic process of a routine character — for which in the deductive sciences the Gödelian process is an example — could likewise be carried out automatically. A routine game of chess can be played automatically by a machine, and indeed, all arts can be performed automatically to the extent to which the rules of the art can be specified. (POLANYI, 1974 [1958], p. 261).

From Polanyi's account, I find two key things for us to learn. First, in Turing's own view, the concept of the universal machine (as I quoted him saying above, "capable of turning itself into any other machine") is connected, indeed, with his 1938 paper (§A.2.2). (It was connected from

the point of view of Polanyi's reminiscence, and I doubt that Polanyi could have made this connection himself if not by recalling Turing's own seminar presentation.) Second, *as of late October 1949 Turing was still referring to the game of chess as intellectual task to illustrate and test for machine intelligence*. Also, from the minute notes of the seminar that survived to us, no reference is made at all about any notion of test for machine intelligence nor to intellectual tasks that would enable it — not from Turing nor from Jefferson either. It seems unlikely that such an important topic could have been part of the discussion and yet been left out of the minute notes. In any case, we see that Polanyi had classed chess as an art that "can be performed automatically" because its rules "can be specified." So Turing had just seen his reference to machine chess-playing to be essentially unimpressive to philosophers. We can now revisit the question left unanswered in the last section: why did Turing make the move of replacing chess by conversational question-answering as intellectual task to test for machine intelligence? Or, why did Turing made the move from physical to thought experiment? The answer shall be straightforward at this point. Physical experiment on the best intellectual task that was feasible at the time would yet not be enough. And indeed, as Robin Gandy's story runs:

> [Turing's 1950 paper] was intended not so much as a penetrating contribution to philosophy but as propaganda. Turing thought the time had come for philosophers and mathematicians and scientists to take seriously the fact that computers were not merely calculating engines but were capable of behaviour which must be accounted as intelligent; he sought to persuade people that this was so. (GANDY, 1996, p. 125)

Turing had to trade his as yet preferred empirical feasibility for appeal, in particular, the field of games for "the learning of languages," which "would be the most impressive, since it is the most human of these activities" (2004 [1948], p. 421). Now, if Polanyi had that influence as of late October 1949, what about Jefferson who is the one Turing actually challenged in his 1950 paper?

Turing provoked by Geoffrey Jefferson

Very recent evidence suggests that another edition of the seminar took place in late 1949. Jonathan Swinton located a Christmas Eve postcard sent to cybernetician Warren McCulloch then in Chicago by a Jules Bogue, an industrialist in the chemical sector and neighbor of Max Newman that found his way into the meeting:

> I wish you [McCulloch] had been with us a few days ago we had an amusing evening discussion with Thuring [*sic*], Wiliams [*sic*], Max Newman, Polanyi, Jefferson, JZ Young & myself. An electronic analyser and a digital computer (universal type) might have sorted the arguments out a bit. (BOGUE, 1949)

In this note appears the name of Professor of Electrical Engineering F. C. Williams (in charge of the Manchester "Baby" computer). Also, some chaos was noted in the arguments during the discussion which may remind us of Turing's motivation to propose the imitation game "as a

basis for discussion" and "criterion for thinking" possibly, as we have seen, after having read Russell's suggestion of the function of the dialectic method in correcting logical errors (§3.3).

Now, I have observed that this finding of Swinton's correlates with what Jefferson related as I quoted before (§1.2). He described an event when Turing would have come to his house to talk to Professor J.Z. Young and him after a meeting in the Philosophy Department. The key information that Jefferson gave was that Turing went there and left on his bicycle "through the winter rain." So, if we take Jefferson's word at face value, that meeting cannot have taken place in late October 1949 (in the fall), and must have been held in late December (in the winter) near Christmas eve. So, it must have been in this December meeting (extended into late night at Jefferson's house) that Turing and Jefferson had some of their most lively exchanges. These must have drawn Turing's attention to Jefferson's Lister Oration, as discussed next.

In fact, in his 1950 paper, not only have Turing cited Jefferson and quoted an excerpt from his. Turing also possessed and annotated an offprint of Jefferson's Lister Oration text (1949a) at the time he was writing his own in early 1950. The offprint of Jefferson's text that was in possession of Turing has been delivered to the King's College Archive at Cambridge University after Turing's death, and the Archive's catalog entry describes it as having "annotations by AMT (Alan Turing)." It was delivered in an envelope also containing the offprint of another and partly related article by Jefferson (1949b).[2]Darren Abramson drew attention to that in (2011). He located two heavy markings in Turing's offprint of Jefferson's Lister Oration (1949a), which gives material evidence that Turing read and annotated Jefferson's text. Altogether I submit that Turing must have done that most likely in the turn of 1949 to 1950, just when he was writing his seminal 1950 paper. (His paper was already accepted by *Mind* and in press in the summer of 1950, so it must have been submitted between late 1949 and early 1950, cf. §A.4.3). Now, in light of all that information, we can narrow down the question on Turing's change of mind: what challenge may have Jefferson posed to Turing that provoked him?

Geoffrey Jefferson (1886–1961), then Professor of Neurosurgery at the University of Manchester and Fellow of the Royal Society since 1947 (WALSHE, 1961), had given on 9 June 1949 in London his Lister Oration — provocatively entitled "The mind of mechanical man" — in virtue of having received the Lister Medal for 1948 by the Royal College of Surgeons of England in recognition of distinguished contributions to surgical science. Days later on 25 June 1949, Jefferson's memorial oration would appear published as an article (1949a). In picking out such a theme for his Lister Oration, Jefferson was motivated and informed by the research and development projects to build modern computing machines (notably the project to build the Manchester "Baby" machine hosted at his own university, cf. §A.3.8), as well as by Wiener's *Cybernetics* published in October 1948 (§A.3.6), as he himself said:

> I have to rely upon and gratefully acknowledge the assistance of Professor F. C. Williams, professor of electro-technics in my own university, and the

[2] Both offprints of Jefferson's (1949a, 1949b) are referenced AMT/B/44 in the King's College Archive's catalog.

> information gleaned from Dr. Wiener, of Boston, in his entertaining book on the new science that he has christened "Cybernetics" (1948).
> (JEFFERSON, 1949a, p. 1108)

The note on *Cybernetics* as an "entertaining book" should not look much ambiguous, as Jefferson had let right in the beginning of his article a caveat:

> No better example could be found of man's characteristic desire for knowledge beyond, and far beyond, the limits of the authentic scientific discoveries of his own day than his wish to understand in complete detail the relationship between brain and mind — the one so finite, the other so amorphous and elusive. [...] We feel perhaps that we are being pushed, gently not roughly pushed, to accept the great likeness between the actions of electronic machines and those of the nervous system. (JEFFERSON, 1949a, p. 1105)

Jefferson may have been aware of Turing already as early as of 9 June 1949 when he gave his memorial lecture, for, as we have already seen (§A.3.6), in the *Cybernetics* Wiener cited Turing multiple times and testified on Turing's leadership on the topic of intelligent machines. For instance, Wiener also cited the engineer F. C. Williams (1965 [1948], p. 122-3). And Jefferson looked up to Williams for learning about the new machines. He reported (cf. above) to have relied upon Williams, who in fact was building the Manchester computer in a (uneasy) collaboration with Turing and Newman (§A.3.8). Jack Copeland reports (cf. 2011, p. 29) he was told by Geoff Tootill (who alongside Tom Kilburn helped Williams to build the "Baby"): "Williams, Kilburn and I [...] disliked [the term 'memory'], incidentally, as encouraging the anthropomorphic concept of 'machines that think'." Not surprisingly, Jefferson seems to have thought highly of Williams:

> To be just, nothing more than analogy is claimed by most of their constructors (some, like Professor Williams, do not go so far even as that), but there is a grave danger that those not so well informed will go to great lengths of fantasy.
> (JEFFERSON, 1949a, p. 1108)

That Wiener and Turing's views in general were seen as outrageous by brain-expert Jefferson, one can get an idea from a book review of Wiener's *Cybernetics* given on 23 February 1949 by a John Thurston writing for *The Saturday Review* (1949). The review was entitled "Devaluing the human brain," and run: "[i]t appears impossible for anyone seriously interested in our civilization to ignore this book," and "overlook cybernetics and its tremendous, even terrifying, implications" (p. 24). Indeed, Jefferson was an Englishman seriously interested in civilization. He took note not only of Wiener's claims but also about the existence of the "Baby" at his own university.

A reporter from *The Times* covered Jefferson's lecture on 9 June 1949 in London. He/she noted Jefferson's strong observation about computers and sonnets, which appeared quoted in his next-day headline "No mind for a mechanical man" 1949b).

Jefferson had required machines to excel at sonnet-writing "because of thoughts and emotions felt," among other demands, in order to accept that "machine equals brain" (p. 1110). His words would also appear quoted by Turing in his (1950) paper over a year later. (I postpone

quoting Jefferson's demands to §3.7 because of its importance to my analysis of the critical function of the imitation game. For a detailed view of Jefferson's demands from a historical point of view, the reader may refer to §A.4.1.)

The reporter from *The Times* must have noticed Jefferson's mention of the Manchester computer (the "Baby") as well, for he/she set out to get in contact with Computing Laboratory in order to get a reply to Jefferson's comments from the computer experts. It turns out, as we know from a letter of Lyn Irvine (Newman's wife) collected by Newman's son William (2012, p. 40), that Newman, who was the director of the laboratory and responsible for the project (§A.3.8), was returning from a trip to Belfast (Northern Ireland) that day, and Turing, as it seems, was the one who was available to respond by a phone interview. Turing, making use of his sense of humor and fine irony (§1.3), responded sharply to Jefferson's statements. The next day (11 June 1949) Turing was thus quoted in *The Times* under headline "Calculus to Sonnet:"

> Mr. Turing said yesterday: "This is only a foretaste of what is to come, and only the shadow of what is going to be. We have to have some experience with the machine before we really know its capabilities. It may take years before we settle down to the new possibilities, but I do not see why it should not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms".
>
> "I do not think you can even draw the line about sonnets, though the comparison is perhaps a little bit unfair because a sonnet written by a machine will be better appreciated by another machine".
>
> Mr. Turing added that the University was really interested in the investigation of the possibilities of machines for their own sake. Their research would be directed to finding the degree of intellectual activity of which a machine was capable, and to what extent it could think for itself. News of the experiments was disclosed by Professor Jefferson in the Lister Oration reported in *The Times* yesterday. (TIMES, 1949a)

From this date on the Jefferson-Turing controversy was established. For context and more details of its development in the public domain, the reader may refer to §A.4.1. From the point of view of my analysis of Turing's imitation tests, Jefferson's Lister Oration itself, which as we have just seen has been read and annotated by Turing, will be the core source indeed.

Jefferson's Oration (1949a) was a daunting critique of the analogy between the new electronic computing machines and the human brain. Essentially, Jefferson condemned the idea that machines could think. He complained that "[w]hen we hear it said that wireless valves think, we may despair of language" (p. 1110). There are six key elements of his text (1949a) which I shall elaborate on in connection with Turing's imitation game. First is Jefferson's exposition of René Descartes' two "very certain means" to distinguish men from machines and animals (p. 1106). His account of Descartes is to the best of my knowledge absolutely precise and correct. Second, Jefferson contended that "sex hormones" are a distinguishing feature of the behavior of "animals" and "men," as opposed to "modern automata" (p. 1107). Third, he argued that the nervous impulse is a continuous signal, and not really an electrical but an electro-chemical

phenomenon. This, for Jefferson, was prohibitive of a strong analogy with computing machines and the phenomena of electronic information storage and processing (p. 1107-8). Fourth, along the same lines of Hartree, Jefferson wrote that "[i]t can be urged, and is cogent argument against the machine, that it can answer only problems given to it, and, furthermore, that the method it employs is one prearranged by its operator." He completed: "[t]he 'facilities' are provided and can be arranged in any order by 'programming' without rebuilding" (p. 1109). Fifth is Jefferson's articulation of the relationship between speech and (conceptual) thinking, which was for him where "there is the sudden and mysterious leap from the highest animal to man" (p. 1109). Sixth comprises, as we have just seen, Jefferson's specific demands to accept that *machine equals brain*, namely, in short, machines should be able to write a sonnet or a concerto because of thoughts and emotions felt (p. 1110). These six topics remarked by Jefferson correspond strikingly to elements of Turing's imitation game.

Overall, I claim, Turing's direct and indirect discussion with these three thinkers, Hartree, Polanyi and Jefferson, is key for any exegesis of Turing's 1950 paper in general, and to an understanding of the conceptual problems he tried to solve in his "imitation tests" in particular. Pressed to respond to the non-trivial challenges posed by them, I suggest, Turing made his crucial 1949 move. He sacrificed his preference for convenience and empirical feasibility and opted for appeal instead. He traded the possibility of showing machine intelligence in a physical experiment based on chess-playing in his own lifetime for arguing about it in a thought experiment based on conversational question-answering which he knew very well that could only be convertible to physical experiments in the future.

3.6 The inner structure of the imitation game

I will start my construction of Turing's imitation game as a thought experiment by presenting its source materials and inner structure. The former are the main elements that Turing drained from cultural and intellectual history and used in the design of the imitation game. The identification of these materials will help explain the latter, which is key to identify the functions of Turing's thought experiment and suggest interpretive boundaries on Turing's proposal.

Ernst Mach established most notably in (1976 [1897]) the use of term *Gedankenexperiment*, which was eternized by his reader Albert Einstein and later translated to "thought experiment." Now, one way of differentiating thought experiments from real or actual experiments is to understand that they may accomplish their epistemic goal — to solve one or more complex conceptual problems, often within some controversy — by means of their design, not their execution. (This is suggested, for instance, by Roy Sorensen, 1992, p. 6.) In the design of his imitation, I found, Turing addressed important problems. There is a large literature on thought experiments in the philosophy of science, which I draw upon but will not review here. Some contemporary references are Jim Brown (1991), Roy Sorensen (1992), and the Norton-

Brown controversy as edited by Christopher Hitchcock (2004). I shall cite others as we go along. According to Mach's classical analysis (1976 [1897]), thought experiments are sourced in quasi-sensory information such as memories and combinations of memories of sense elements (p. 137). Turing seems to have been a listener of BBC radio, and his crucial 1949 exchanges also seem to have caught his imagination. I shall claim that the imitation game has been designed as a blend of the materials presented next.

The source materials

Twenty Questions was a radio parlor game originated in the United States in the nineteenth century whose popularity escalated in the late 1940s, when then a British version started to be run and broadcasted by BBC. (Specifically, the BBC aired a version on radio from 28 February 1947 to 1976 with TV specials airing in 1947 and 1948.) On radio, the subject to be guessed was revealed to the audience by a "mystery voice." The players were allowed to ask up to twenty questions about a mystery object in their quest to identify it. The only clue supplied was whether the item was of animal, vegetable or mineral nature. (Note the ontological nature of the clue and its focus on species.) The program had been a staple on radio since 1946.[3] It is also interesting to note, in terms of the logical structure of the game, that careful selection of questions can greatly improve the odds for the questioner to win the game; and in terms of the social and historical culture, that conservative gender-based questions, *e.g.*, "Can I give it to my mother-in-law?" or "Can I do it to my wife?," were a high moment in these shows. (Note the suggestion of the gender issue, not only from the point of view of Turing as a homossexual living in postwar Britain, but also as an important element in the asking for a clue and in the guessing itself.)

As if the similarity between structural elements of this radio parlor game and Turing's imitation game were not enough, there is also (primary-source) direct mention of the former by Turing himself in passing (1950, p. 457) and a (secondary-source) guess by his contemporary witness Wolfe Mays (1952, p. 148).

We have seen before that Turing referred to "[his] imitation tests" (2004 [1952], p. 503), and that he in fact varied the conditions of his imitation game to produce several variants of it (§3.3). This seems to fit well to Mach's analysis of thought experiments. The way Turing drained *Twenty Questions* as source material from culture into the design of his thought experiment is another piece of evidence in this connection.

Blended with the radio parlor game, furthermore I identify two other pieces of material in which Turing seems to have sourced the design of his 1950 thought experiment. In this case

[3] Cf. IMDb entry <https://www.imdb.com/title/tt0320997/>. For a recording of uncertain date from the late 1940's to the early 1950's of the BBC radio version that must have inspired Turing, the reader may refer to <http://www.youtube.com/watch?v=rSN49JOa4cs>. The show, it is stated, was "chaired by the irascible Gilbert Harding and featuring panelists Joy Adamson, Anona Winn, Jack Train and Kenneth Horne — not forgetting the sepulchral 'mystery voice' of Norman Hackforth. This edition was broadcast from Brighton." These participants match BBC's own listing for its Home Service at 20:30 on 22 June 1950, as available at <http://genome.ch.bbc.co.uk/page/7c84364ebe86449da5c8ce903245fb78>. All accessed on 3 Sep. 2020.

they come not from material culture but directly from the very controversy he was engaged in. Let us recall from (§3.5) Jefferson's Lister Oration and Darren Abramson's related testimony (2011, p. 548-9). Abramson reported to have found in the offprint of Jefferson's text (1949a) that was in possession of Turing *c*. 1949-1950 a heavy line marked on the margin of two core passages, namely, Jefferson's review and citation of Descartes (p. 1106) and Jefferson's demands on machine intelligence (p. 1110) that Turing quoted in his 1950 paper. Regarding the first of these passages, why did Turing mark it? Below I shall quote the passage in full, in three parts:

> Descartes made the point, and a basic one it is, that a parrot repeated only what it had been taught and only a fragment of that; it never uses words to express its own thoughts. If, he goes on to say, on the one hand one had a machine that had the shape and appearance of a monkey or other animal without a reasoning soul (i.e., without a human mind) there would be no means of knowing which was the counterfeit. On the other hand, if there was a machine that appeared to be a man, and *imitated his actions* so far as it would be possible to do so, we should always have two very certain means of recognizing the deceit.
> (JEFFERSON, 1949a, p. 1106, emphasis added).

So Descartes offered to Turing via Jefferson the image of a machine that would resemble a man and "imitate his actions" as closely as possible. Jefferson quoted Descartes, as known, from Part V of the *Discourse*, firstly published in (1985 [1637]). Descartes have indeed offered this image just as Jefferson phrased, cf. Robert Stoothoff's translation (*Ibid.*, p. 139). We shall see later (§3.7) that this could hardly have passed unnoticed to Turing, who in February 1947 had literally referred to the capacity of his "universal machine" to "imitate" any other machine. Jefferson thus resumed with the description of the "two very certain means" proposed by Descartes:

> First, the machine could not use words as we do to declare our thoughts to others. Secondly, although like some animals they might show more industry than we do, and do some things better than we, yet they would act without knowledge of what they were about simply by the arrangement of their organs, their mechanisms, each particularly designed for each particular action (cp. Karel Capek's Robots). (JEFFERSON, 1949a, p. 1106).

And Jefferson reproduced Descartes's confidence about the prospects one should expect:

> Descartes concluded: 'From which it comes that it is morally impossible that there be enough diversity in a machine for it to be able to act in all the occurrences of life in the same way that our reason would cause us to act. By these means we can recognize the difference between man and beasts.' He could even conceive a machine that might speak and, if touched in one spot, might ask what one wanted — if touched in another that it would cry out that it hurt, and similar things. But he could not conceive of an automaton of sufficient diversity to respond to the sense of all that could be said in its presence. It would fail because it had no mind. (JEFFERSON, 1949a, p. 1106).

A key question of the passage is whether or not there could be "sufficient diversity" in a machine for it to be able to imitate a man. Descartes, as put by Jefferson at the end of his citation above,

"could not conceive of an automaton of sufficient diversity to respond to the sense of all that could be said in its presence." For him, it was not practically possible for machines to have as much diversity of behavior, verbal or otherwise, as we humans do. But in fact, the reference of machine that Descartes considered were the hydraulic automata of the Renaissance gardens (WERRETT, 2001). Because of the lack of enough diversity in such machines, Descartes could not imagine any machine that could effectively imitate us, and think. Were we to stop here, the similarity between the source material derived from Descartes's 1637 text and structural elements of Turing's imitation game is itself significant. But there is more. I shall now take a moment to introduce the third piece of source material of Turing's 1950 thought experiment, which also comes from Jefferson's 1949 text.

Jefferson, while subscribing to Descartes's (1637) "two very certain means," attached a different (1949) thesis to them. Like Descartes, Jefferson held that machines can't think. But Jefferson, professor of neurosurgery, had a different reason. It is the biochemical substrate of the brain that produces "mental process" and "thoughts," he held:

> Descartes solved the difficulty by making mind supernatural, placing an immaterial mind independent of organism in the pineal. [...] We may well doubt to-day whether a supernatural agency is the basis of mental process. But it was doubted in Lister's time. In 1870 T. H. Huxley reluctantly concluded: "[...] the thoughts to which I am now giving utterance, and your thoughts regarding them, are the expression of molecular changes in the matter of life which is the source of our other vital phenomena." (JEFFERSON, 1949a, p. 1106-7).

So Jefferson approximated himself to Huxley's reduction of mind to brain. On the one hand, he echoed to the two cartesian means for distinguishing machines from men. On the other hand, he differed from Descartes with regard to scientific and/or philosophical assumptions on causes. Given the particular road that Jefferson took to establish the superiority of men over machines, he seems to have felt himself compelled to concede "conscious mental processes" to animals:

> It is the infinite variety of the behaviour of the world of animals that confuses us. The stage is too vast, the cast too numerous, the qualities of their performances too varied. We should not show any hesitation in attributing conscious mental processes to animals to-day. (JEFFERSON, 1949a, p. 1107).

But then Jefferson went on to set things straight again and differentiate "the highest animal" from "man" (*sic.*). He returned to Descartes and suggested speech or the use of language as evidence of the highest intellectual faculties:

> Granted that much that goes on in our heads is wordless (for if it is not, then we must concede words, an internal vocabulary, to animals), we certainly require words for conceptual thinking as well as for expression. It is here that there is the sudden and mysterious leap from the highest animal to man, and it is in the speech areas of the dominant hemisphere [of the brain] rather than in the pineal that Descartes should have put the soul, the highest intellectual faculties. (JEFFERSON, 1949a, p. 1109)

In sum, one could say, Jefferson purported to update Descartes's beast-machine thesis. He tried to demarcate machines out of the realm of thinking on account of their lack of a biochemical brain and demarcate "man" as above "the highest animal" on account of a "sudden and mysterious leap" operated by evolution. We shall now be ready to see the third source material that Turing blended, I suggest, into the design of his imitation game.

To argue for his point, Jefferson referred to "sex hormones" as a distinguishing feature of the behavior of "animals" and "men," as opposed to "modern automata" (p. 1107). Specifically, he cited cybernetician Grey Walter's iconic electromechanical tortoise, and wrote:

> In a favourable situation the behaviour of such a toy could appear to be very lifelike — so much so that a good demonstrator might cause the credulous to exclaim "This is indeed a tortoise." I imagine, however, that another tortoise would quickly find it a puzzling companion and a disappointing mate. (JEFFERSON, 1949a, p. 1107)

Shortly after in the same page, Jefferson remarked that "neither animals nor men can be explained by studying nervous mechanics in isolation, so complicated are they by endocrines, so coloured is thought by emotion." He completed: "[s]ex hormones introduce peculiarities of behaviour often as inexplicable as they are impressive." Indeed, Jefferson suggested that machines could not exhibit enough peculiarities of behavior to be able to imitate the actions of animals or "men" because they have no sex hormones. A machine would give itself away and be found to be "a puzzling companion and a disappointing mate." Further on in his text, when outlining his demands that went quoted in *The Times*, Jefferson posed that:

> No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, *be charmed by sex*, be angry or miserable when it cannot get what it wants. (JEFFERSON, 1949a, p. 1110, emphasis added)

In his (1950) formulation of objection "(5) arguments from various disabilities," Turing seems to respond to Jefferson's claims in general, including the claim that no mechanism could be "charmed by sex" in particular. He rebutted various claims that one cannot make machines "to do X," and in particular, to "fall in love" and "make some one fall in love with it" (p. 447). Now, as known, gender imitation is an element of Turing's historical imitation game. Although it has been eliminated from Turing's test by some commentators, it does has been acknowledged as an important element of it by others. And yet, to my knowledge, the connection with Jefferson's passage has never been noted. Biographer Andrew Hodges (2012 [1983]), for example, even came to write that Turing's "sexual guessing game" was "in fact a a red herring, and one of the few passages of the paper that was not expressed with perfect lucidity" (p. 415). Hodges thus continued his struggle with Turing's text: "[t]he whole point of this game was that a successful imitation of a woman's responses by a man would not prove anything." But "[g]ender," he argued again buying Jefferson's views (cf. §1.1), "depended on facts which were *not* reducible

to sequences of symbols" (no emphasis added). Now, whatever scientific and philosophical analysis one may propose in this connection, my point here is nothing but to show that gender imitation — which is in point of fact present in at least the first of the various experimental questions that Turing asked — turns out to find a corresponding source material in Jefferson's 1949 text. Jefferson held in that occasion, as he would say more explicitly to Turing's face in 1952, that "[m]an is essentially a chemical machine" (cf. §A.4.6, J14; or 2004 [1952], p. 502). Indeed, Jefferson championed the assumption that a machine could not pass to be an animal or man('s mate) because it lacks sex hormones. Turing disagreed. And the inclusion of gender imitation as part of his test for machine intelligence, I hold, encoded a brilliant ironic response to Jefferson. Right or wrong, for the scope of machine capabilities is yet undecided, we will be able to appraise the scientific and philosophical value of Turing's response.

Altogether, I claim that the imitation game has been designed as a blend of essentially these pieces of source material. It has drawn on the radio parlor game aired by BBC, on Jefferson's account of Descartes's two "very certain means" to distinguish "men" from machines and other animals, and on Jefferson's image about sex hormones. I shall now try to show this in detail.

The inner structure

Turing considered the question (Q) whether machines can think and thus defined "[t]he new form of the problem," which he described in terms of the "imitation game:"

> It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:
>
> C: Will X please tell me the length of his or her hair?
>
> Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be
>
> 'My hair is shingled, and the longest strands are about nine inches long.'
> (TURING, 1950, p. 433-4)

We see that the imitation of gender is the focus of Turing's example of interrogation, indeed. He then proceeded to define some settings that he must have felt were important in its design. He ruled out from the game sensorial signs such as tones of voice. The ideal arrangement, Turing defined, is to have a teleprinter communicating between the two rooms:

> In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as 'I am the woman, don't listen to

> him!' to her answers, but it will avail nothing as the man can make similar remarks. (TURING, 1950, p. 433-4)

So in this standard form of the game, there are three players A, B and C with fixed goals, just as in Turing's 1948 unnamed game based on chess-playing (§3.4). Initially, these roles are played by "a man (A), a woman (B), and an interrogator (C) who may be of either sex" (p. 434). Let us now take the pains to systematically recollect the various experimental questions $Q' \ldots Q''''$ that Turing asked based on his imitation game to replace the original question Q:

(Q'). "'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" (p. 434). Note that this is a *computer-imitates-woman* version of the game.

(Q''). "There are [...] digital computers in working order, and it may be asked, 'Why not try the experiment straight away? [...]' The short answer is that we are not asking whether all digital computers would do well in the game nor whether the computers at present available would do well, but whether there are imaginable computers which would do well" (p. 436).

(Q'''). "'Let us fix our attention on one particular digital computer C. Is it true that by modifying this computer to have an adequate storage, suitably increasing its speed of action, and providing it with an appropriate programme, C can be made to play satisfactorily the part of A in the imitation game, the part of B being taken by a man?"' (p. 442). Note that this is a *computer-imitates-man* version of the game.

(Q''''). Whether "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" (p. 442).

Turing stated belief that the answer to the fourth variant (Q'''') of the question would be yes. Note that in the first version (Q') Turing proposed a decision rule about whether or not a machine passes the *machine-imitates-woman* test. It shall depend on whether or not "the interrogator decide[s] wrongly as often when the game is played like this as he does when the game is played between a man and a woman." He implicitly suggested that this decision rule was kept in the second and third versions (Q'', Q'''). This does not seem to hold for the fourth one. In this case, however, he was seemingly not suggesting a new decision rule for the game as an experiment but rather making a rough prediction about it.

Figure 3 shows a pictorial representation of Turing's imitation game having a digital computer as one of the players. Turing gave further "specimen questions and answers" for illustration and thus extended it from his initial gender-centered example:

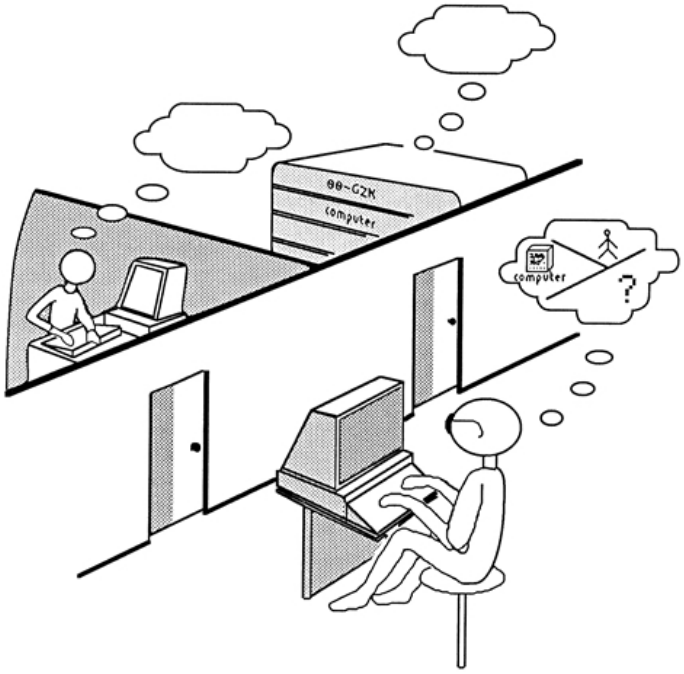

Figure 3 – Illustration of Turing's imitation game or test. Drawing by Ann Witbrock, copyright of B. Jack Copeland (1993). Reproduced with permission.

> 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 K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?
> A: (After a pause of 15 seconds) R-R8 mate. (TURING, 1950, p. 434-5)

I want to draw attention to what goes on in this example. Note that the first question asks for poetry, and the machine shies away from it. Then an arithmetical question is made that could unveil the machine (playing A), say, if it responded too fast and accurate. But the machine takes a pause and makes a mistake (responds 105621 instead of 105721). The interrogator shifts the conversation to the game of chess. And in this branch of thought or field the machine thrives. So in order to give a prompt illustration of the kind of *viva-voce* examination he had in mind Turing showcased a machine that is bad at poetry and excellent at chess. This shows that Turing was confident about making a machine to perform well in chess-playing, while poetry he seems to have had as a more challenging intellectual task. And yet later on in the paper, in a dialogue with Jefferson relative to the "argument from consciousness" against the possibility of machine thinking, he foresaw a (future) machine that had poetic skills. I shall come back to this point later on (§3.7). Overall, in this sample of questions and answers Turing seems to have been concerned with showing, and in any case has shown, that conversational question-answering, in fact, allows

one to interrogate the players of the game in almost any intellectual field one likes. Turing argued just that (1950): "[t]he question and answer method seems to be suitable for introducing almost any one of the fields of human endeavour that we wish to include" (p. 435). In particular and perhaps most interestingly, Turing showed that unrestricted conversation as an intellectual task even subsumes the game of chess, which was Turing's first choice when he was more concerned with a physical experiment and less with a thought experiment for machine intelligence.

This leads us to the actual analysis of the structure of the imitation game, which I interpret to be a blend of (i) the radio parlor game aired by BBC, (ii) Jefferson's account of Descartes's two "very certain means" to distinguish "men" from machines and other animals, and (iii) Jefferson's image about sex hormones. Turing combined these materials, as we will now see.

In (1964a, p. 198), Keith Gunderson called the two cartesian means to distinguish animals and machines from men the *language test* (the first appearing in Jefferson's citation in §3.6) and the *action test* (the second one). The term "test" never actually occurs in, say, Robert Stoothoff's translation of Part V of the Descartes's *Discours de la méthode & Essais* (1985), henceforth the *Discourse*. While Gunderson himself seems to admit to be engaging in some anachronism (cf. 1964a, p. 198), I rather found in Part VI of the *Discourse* (p. 143) a relevant translation note on the term "observations," suggesting Descartes's uses of French *expériences* as "a term which [he] often uses when talking of scientific observations, and which sometimes comes close to mean the same as "experiments" in the modern sense (its root being derived from Lat. *experior*, 'to test')." Now, we know that Descartes was committed to scientific observations on (i) the performances of animals such as "magpies and parrots" which "can utter words as we do" (1985, p. 140) — in connection with his so-called "language test" —; and (ii) the performances of the hydraulic automata of the Renaissance gardens (WERRETT, 2001) which could "move in many different ways" in reaction to the "mere presence" of external objects (1985, p. 101) — in connection with his so-called "action test." So Gunderson's neologism over the two cartesian means does not seem to be unsound and I shall use it.

The relative roles of the language and the action tests received heterogenous interpretations in the literature. Gunderson understood that "the language test [...] is exactly the same *type* of test as the action test, with the exception that it subsumes a more specific (though still very broad) range of activities" (1964a, p. 199, no emphasis added). Others followed that interpretation. For example, Gerald Erion (2001) took the linguistic test to be more restricted than the non-linguistic one on account that the latter (but not really the former) is a test for commonsense knowledge. Erion so concluded that Turing's 1950 test is weaker than Descartes's, as it would have left uncovered some situations that are covered by Descartes' (purportedly harder) action test. As a reading of Descartes, nonetheless, I say that Erion's interpretation pays little attention to the relative importance of the two tests according to Descartes himself, for whom the action test was rather secondary. This can be seen from Descartes's order of exposition of the two tests in the *Discourse* (main source) but also most explicitly in his (1991 [1649]) letter to Henry More,

twelve years after the main source, where he pointed out that "speech is the only certain sign of thought hidden in a body" (p. 366). For Descartes, the language test had precedence over the action test. In fact, the language test offers a convenient way to emulate a very broad scope of situated actions. Turing observed this himself, as we have seen from his comment about the "question and answer method," which is "suitable for introducing almost any one of the fields of human endeavour" (p. 435).

Now, there should be little doubt at this point that Turing's 1950 imitation game addresses Descartes's language test, and not his action test. Turing justified that "[t]he question and answer method" has enough breadth of scope to allow for "for introducing almost any one of the fields of human endeavour that we wish to include." That is because, as we have seen, it allows for introducing in conversation whatever action situation one can talk about. But not only Turing did argue for the convenience of what we can associate with the language test. He also argued against the action test, as if fixing an important feature of Descartes's characterization, namely, that any test for machine (or animal) intelligence must not be prone to confirmation bias based on "physical" capacities. Recall from Jefferson's citation of Descartes the passage: "[h]e could even conceive a machine that might speak and, if touched in one spot, might ask what one wanted — if touched in another that it would cry out that it hurt, and similar things." And Turing wrote:

> The new problem has the advantage of drawing a fairly sharp line between the physical and the intellectual capacities of a man. No engineer or chemist claims to be able to produce a material which is indistinguishable from the human skin. It is possible that at some time this might be done, but even supposing this invention available we should feel there was little point in trying to make a 'thinking machine' more human by dressing it up in such artificial flesh. The form in which we have set the problem reflects this fact in the condition which prevents the interrogator from *seeing or touching the other competitors, or hearing their voices*. (TURING, 1950, p. 434, emphasis added)

Turing wrote it almost as if he was justifying why he came up with a modification of Descartes's 1637 proposal as cited by Jefferson in 1949. He completed:

> We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane. The conditions of our game make these disabilities irrelevant.
> (TURING, 1950, p. 435)

Now, let us recall the radio parlor game. A core feature of it is that one has to guess the ontological kind of an unknown entity without being able to hear its voice, to see or to touch it. The only way to get to know more and more about it was to inquire by making questions. Turing's imagination seems to have been caught by the radio. He introduced in the imitation game an arrangement for blind communication between the contestants and the interrogator.

This was not the only adaptation that Turing operated in Descartes's language test. In Descartes's description, there are only the one contestant entity — animal, machine or human —

and the human interrogator who inquiries it. But besides introducing in the imitation game an arrangement for blind communication, Turing also introduced an arrangement for a third player B supposed to help the interrogator making the right decision. Now, let us recall Jefferson's argument about the influence of sex hormones in the production of peculiarities of behavior in "animals" and "men." It turns out that Jefferson offered the image of an electromechanical tortoise that is put side by side with an actual tortoise. Suppose we consider something analogous for humans, say, as Jefferson suggested in the title of his Lister Oration, a man side by side with a "mechanical man." Without being able to see, touch or hear the two, would a human inquirer be able to tell them correctly? Or would the interrogator, also as put by Jefferson in the context of his critique of Grey Walter's electromechanical tortoise, "quickly find it a puzzling companion and a disappointing mate"? If the thesis that sex hormones are crucial to produce specific and complex behavior was also at stake, then Descartes's language test had to be extended to account for gender behavior, indeed. General human intelligence would not be enough. Gender-specific intelligence — purportedly an even harder challenge — would be required. These two adaptations, one derived from the radio parlor game and the other from Jefferson's own text, I think, seem to extend Descartes's language test to match strikingly the arrangements introduced by Turing into his imitation-game thought experiment.

In sum, to offer a schematic view of the inner structure of the imitation game by paying attention to its historical sources, we may consider these various schemes:

- There is a human interrogator (player C) that questions a contestant entity (player A), which gives answers so that a conversation may unfold. Player C can see, touch and hear the inquired entity. Player A may rely on trickery to deceive C about its true ontological kind ("man," "animal" or "automaton"), while C may rely on the unrestricted scope of the questions that can be made to distinguish A correctly. The indistinguishability from a human being in performing an intellectual task is taken as epistemological criterion for thinking. This specific scheme comes from Descartes's 1637 *Discourse* and has been reproduced by Jefferson in 1949. Because of its choice for conversational question-answering as the preferred intellectual task to be performed by the contestant entity A in its pretence of showcasing intelligence or thinking, it can be called a language test.

- An arrangement can be introduced in the language test to make the communication blind between the human interrogator (C) and the contestant entity (A). The ideal arrangement is to use different rooms for each player and convey the conversational question-answering by teletyping. This scheme has been proposed by Turing in 1950, seemingly by inspiration from *Twenty Questions*, a radio parlor game aired by BBC at the time, in order to draw a fairly sharp line between the physical and the intellectual capacities of the inquired entity, and then avoid biases in the evaluation of its intelligence. This is a blind language test.

- A third player (B) can be introduced in the blind language test to help the human inter-

> rogator (player C) to make the correct judgement. The best strategy for player B is to give truthful answers. The presence of this additional player B side by side player A (whose pretence is to try to pass to be an ontological kind different than its own) is to serve as a baseline model of the gender performance (of either a woman or a man, depending on the version of the game that is chosen to be played) in a unrestricted conversation. If the contestant entity can imitate well the chosen gender, then it will showcase not only human intelligence in general but also the "peculiarities of behaviour" that would be rendered by specific "sex hormones," if this makes any sense at all for an evaluation of intelligence. Last but not least, the presence of a human contestant sharply shifts the burden of proof towards balancing it with the human side. This scheme has also been proposed by Turing in 1950, seemingly to respond to a thesis laid out by Jefferson in his 1949 text. This is a blind language test but also a gender test in particular and a species test in general, for the imitation of a woman or a man entails (and embodies) the imitation of the human.

It can be observed that Turing modified Descartes's language test to address additional issues and to prevent confirmation bias against the machines. I shall refer to these three schemes as (i) Descartes's language test, (ii) the blind communication arrangements, and (iii) the presence of a human third player. They are each essential to make up the historical imitation game as presented by Turing in his 1950 paper. They are each functional to represent Turing's view of intelligence.

Overall, the identification of the source materials and Turing's use of them in the design of the imitation game shall help us understand its dual function as a thought experiment. I have remarked in the beginning of this section that thought experiments may accomplish their epistemic goal by means of their design and not by their execution. We shall observe this in the case of Turing's imitation game by studying its functions.

3.7 The dual function of the imitation game

Thought experiments in science have been classified in terms of their functions. Let us take a moment to get acquainted with the terminology. A taxonomy was offered by Jim Brown (1991). He referred to Karl Popper's discussion of "the *critical* and the *heuristic* uses of thought experiments, which corresponds roughly to my destructive and constructive types" (p. 34, no emphasis added). Brown added that given Popper's views on theory confirmation (as known, Popper tried to rule any such notion out of the empirical sciences), however, the similarity of their views was limited. I subscribe to Brown's observation about Popper's views, and yet shall follow Popper's neater terminology of "critical" and "heuristic" uses or functions of thought experiment (2002 [1959], p. 465). Let us see how Popper described them, starting with the critical function:

> One of the most important imaginary experiments in the history of natural philosophy, and one of the simplest and most ingenious arguments in the history of rational thought about our universe, is contained in Galileo's criticism of

> Aristotle's theory of motion. It disproves the Aristotelian supposition that the natural velocity of a heavier body is greater than that of a lighter body.
>
> I see in Galileo's imaginary experiment a perfect model for the best use of imaginary experiments. It is the *critical* use.
> (POPPER, 2002 [1959], p. 464-5, no emphasis added)

Let us say, in light of this, that one possible function of a thought experiment is to criticize one or more assumptions about a studied phenomenon. (We see that Popper, like Dennett as we have seen before and even Brown too, suggested that Galileo's thought experiment refuted Aristotle's theory of motion by logic alone. This is however far from obvious, as we shall see later (§3.9). It is important to emphasize, for now, that the critical function of a thought experiment is, of course, *to criticize* (say, assumptions of a rival theory). Going further and requiring "refutation" for its fulfillment may rather engage in a romanticized view of the performance of thought experiments in scientific controversies. In connection with the critical or "destructive" function of a thought experiment, Brown put that it "destroys or at least presents serious problems for a theory, [...being it] anything from a minor tension [...] to a downright contradiction" (1991, p. 34). Also, the reference to "theory" should be deflationary here, as often what is at stake is an assumption that belongs to a view but not necessarily to a full-fledged theory. Let us now move to Popper's description of the heuristic function of thought experiments:

> I do not wish to suggest, however, that there is no other way of using them. There is, especially, a *heuristic* use which is very valuable. [...]
>
> Heuristic imaginary experiments have become particularly important in thermodynamics (Carnot's cycle); and they have lately become somewhat fashionable owing to their use in relativity and in quantum theory. One of the best examples of this kind is Einstein's experiment of the accelerated lift: it illustrates the local equivalence of acceleration and gravity, and it suggests that light rays in a gravitational field may proceed on curved paths. This use is important and legitimate. (POPPER, 2002 [1959], p. 465-6, no emphasis added)

So for Popper the heuristic function "illustrates" and/or "suggests" a property of a studied phenomenon. Brown in turn referred (1991) to this particular function as "mediative," as it "facilitates a conclusion" (p. 36). In short, it shall be clear at this point that thought experiments may have a negative (critical) and a positive (heuristic) function. Brown acknowledged that Galileo's famous thought experiment mentioned above is one of the few that had both of these functions (p. 43). I find both functions in Turing's imitation game as well, as presented next.

The critical function

A thought experiment, as mentioned, can be used to criticize one or more assumptions about a studied phenomenon. Now, the imitation game is an experiment about the phenomenon of thinking or intelligence. So what assumptions did Turing use his thought experiment to criticize?

Turing was confronted by assumptions posed most notably, as we have seen, by Hartree, Polanyi and Jefferson. Some of them he seems to have identified as reasonable given the state

of knowledge back then, most notably, Polanyi's point that chess would not be very impressive to showcase machine intelligence and the Lovelace-Hartree thesis. These Turing discussed positively by means of the heuristic function of the imitation game, as we shall see in the sequel (§3.7). But other assumptions Turing seems to have seen as logical errors that needed to be fixed. He seems to have thought of these as (in my words) *a priori* or metaphysical assumptions about human behavior which, first, violated his plea for "fair play for the machines" (§1.6), and second, could never be checked empirically. These in turn include (i) Polanyi's assumption about the informality of human behavior; Jefferson's assumptions about (ii) raw feels, (iii) sex hormones as key to the variety of animal behavior, and (iv) the incommensurability between the nervous system (continuous) and computing machines (discrete); and (v) Descartes's assumption of a rational soul, and about (vi) the impossibility of really diverse behavior in machines. Turing discussed all of them in his negative dialectics by making a critical use of the imitation game, except for Descartes's assumption (v) which he addressed as the "theological objection" (cf. Figure 1) with no reference to the game. We shall now see them by following the order above.

In his formulation of the eighth objection to machine intelligence, the "argument from informality of behavior" (1950, p. 452-3; Figure 1), Turing discussed whether or not human behavior could be encoded into a set of rules and thus programed into a digital computer. We can associate this assumption with Polanyi, as can be seen from the notes of the October 1949 Manchester seminar (§A.4.2). Turing replied by suggesting that there must be some *a priori* assumption underlying this argument. What makes us think that our behavior is not amenable to rule specification does *not* seem to be the observation of external human behavior. For if the same criterion is applied to the behavior of a machine — that is, if we judge its behavior by the observation of its external behavior alone —, such a set of rules could hardly be found either. He reported to have himself experienced it with the Manchester "Baby" universal computer. This, Turing observed, was in spite of the fact that the behavior of the computer *was* in fact specifiable by rules. So, Turing concluded, whether or not some behavior is specifiable by rules is underdetermined by the observation of behavior alone. In fact, he argued, machine behavior is encoded by rules but perhaps our own behavior as humans is also encoded by rules (laws of nature) even though we do not know them and are not aware of them. Now, note that Turing's response is clearly represented in the imitation game. As we have seen, Turing adapted Descartes's language test by introducing arrangements for blind communication and the presence of a human third player. The interrogator was then supposed to observe the verbal behavior of the human and the machine without knowing from the start who is who. The issue can only be resolved by making questions and guessing out of them. So, by denying the interrogator to know beforehand who is who, the imitation game neutralized the *a priori* (metaphysical) assumption behind the argument from informality of behavior. Let us now move to Jefferson's assumptions.

Jefferson had outlined his demands in order to accept that "machine equals brain:"

> [N]ot until a machine can write a sonnet or a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain — that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or miserable when it cannot get what it wants. (JEFFERSON, 1949a, p. 1110)

This passage was quoted in *The Times* on 10 June 1949. And Turing also quoted it in his 1950 paper in what turned out to be his most explicit reply to Jefferson's demands. He then wrote:

> This argument appears to be a denial of the validity of our test. According to the most extreme form of this view the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking. According to this view the only way to know that a *man* thinks is to be that particular man. (TURING, 1950, p. 446, no emphasis added)

Turing thus rejected the relevance of raw feels or qualia for machine intelligence, for these seemed *not* to be inter-subjectively observable either directly or indirectly. He was not able to neutralize this assumption of Jefferson's within the imitation game design itself. (And this is in spite of whatever Turing thought about the possibility of a machine to actually have raw feels and consciousness. Turing's views on this are discussed in §2.5. My focus here is on the critical function of the imitation game.) However, he *used* the imitation game to show that Jefferson's assumption was *a priori* and in violation of fair play. And represented in his thought experiment he retained an important part of Jefferson's challenge, namely, an idealized sonnet-writing machine which could discuss its own sonnet charmingly. He thus returned to Jefferson the challenge:

> I am sure that Professor Jefferson does not wish to adopt the extreme and solipsist point of view. Probably he would be quite willing to accept the imitation game as a test. The game (with the player B omitted) is frequently used in practice under the name of *viva voce* to discover whether some one really understands something or has 'learnt it parrot fashion'.
> (TURING, 1950, p. 446, no emphasis added)

The machine was interrogated by a judge about a sonnet it would have written:

> Let us listen in to a part of such a *viva voce*:
>
> 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 flay one means a typical

> winter's day, rather than a special one like Christmas.
> (TURING, 1950, p. 446, no emphasis added)

Turing then concluded his response to Jefferson's demands in order to accept that "machine equals brain" by provoking Jefferson:

> And so on. What would Professor Jefferson say if the sonnet-writing machine was able to answer like this in the *viva voce*? I do not know whether he would regard the machine as 'merely artificially signalling' these answers, but if the answers were as satisfactory and sustained as in the above passage I do not think he would describe it as 'an easy contrivance'.
> (TURING, 1950, p. 446-7, no emphasis added)

This was a high moment of Turing's 1950 paper and his illustration of the imitation game.

Note that Jefferson seems to have chosen poetry just for its apparent intractability. So, Turing's shift from physical to thought experiment in between October 1949 and early 1950 as we have seen (§3.5) must have been also to respond to Jefferson's challenge. (Hartree and Polanyi's arguments must have been considered too.) But once Turing had made that move, the sonnet-writer machine seems actually to have been a sensible, if not perfect example to criticize Jefferson's appeal to raw feels as absurd. Turing was seemingly aware of how ahead of his time the sonnet-writer machine was, and he knew that speech was not even close to be feasibly addressed in a test back then. Now, in connection with Turing's reply to the mathematical objection after his experimental turn (§A.2.2), we may also note that the topic of sonnet-writing appear two times in (1950) by means of two different illustrations. In a first example (p. 434), the machine replies to the interrogator by acknowledging its limitation: "I never could write poetry," it says. In the second example which we have just seen, however, it can. Turing this showed that different machines may have different competencies, just as he had told Newman in his (2004 [*c.* 1940]) letter. No particular machine would be able to perform well in all tasks, just as we humans could not. Let us now move to what was in Turing's view Jefferson's second logical error.

Jefferson, as we have seen in depth (§3.6), had presented a view in which the sex hormones of an individual animal (non-human or human) produce a crucial part of the individual's variety of behavior relative to gender (1949a, p. 1107). He offered the image of an electromechanical tortoise that is put side by side with an actual tortoise. A machine in this situation would give itself away and be found to be "a puzzling companion and a disappointing mate." Jefferson, in effect, championed the assumption that a machine could not pass to be an animal or man('s mate) because it lacks sex hormones. Now, Turing had introduced in the imitation game the presence of a gendered human third player B. He then asked whether the machine (player A) would be able to imitate player B — either a woman or a man depending on the version of the game — well enough to deceive the interrogator (player C). So Turing placed the interrogator in the position of having to distinguish between the machine's imitation of a specific human gender and the behavior of a legitimate representative of that gender. And in his positive

discussion, Turing suggested that machines could learn from experience and thus play well the imitation game. So Turing implied that gender-specific behavior can be learned (by a machine) and thus is not determined by sex hormones, as Jefferson claimed. In fact, by introducing gender imitation into the design of the imitation game, Turing rendered a brilliant scheme to neutralize Jefferson's *a priori* assumption about the relation (if any) between the physiology of behavior and gender. It is absolutely astounding to observe that Turing designed his imitation game to challenge Jefferson's thesis as of *c.* early 1950, and two years later he would be imposed by the British State a pseudo-therapy based on sex hormones to convert his homossexual behavior, as we have seen in §1.7. In fact, Turing was *not* converted to heterosexuality after this deplorable treatment (HODGES, 2012 [1983], §8). Rather he was *himself* proof that Jefferson was wrong.

Turing addressed Jefferson's third assumption in his formulation of "the argument from continuity in the nervous system" (1950, p. 451-2). Jefferson had dedicated one section of his Lister Oration to discuss four topics relative to the nervous impulse (in my words): (i) speeds of the electronic computing machine compared to the nervous impulse, (ii) speed of thought, (iii) counter-evidence to the electrical nature of the nerve impulse, and (iv) methods of study and the chemical agencies of the nervous system. This was an important part of his general Lister-Oration argument — which he offered most concisely in his fourteenth intervention in the January 1952 BBC roundtable — "man is essentially a chemical machine" (§A.4.6). In reply, Turing dedicated only a few lines to rebut this one. He acknowledged that "[t]he nervous system is certainly not a discrete-state machine," for "[a] small error in the information about the size of a nervous impulse impinging on a neuron, may make a large difference to the size of the outgoing impulse" (1950, p. 451). He pondered, however, that this does not mean that the behavior of the nervous system cannot be mimicked by a discrete-state system. He argued that the imitation game design neutralizes this ontological difference. Turing considered an example where the interrogator asks the contestants — one is a digital computer, and the other is a differential analyzer (a simpler and yet continuous system) — to answer the value of a transcendental number such as π. The digital computer could imitate the differential analyzer by choosing at random from a probability distribution between values that approximate the true answer (say, 3.1416). So, more generally, we can say, the discrete-state machine may use any mathematical technique to approximate well enough the behavior of the continuous-state machine and yet an external observer (the interrogator) may not be able to distinguish which is which. That is, in the imitation game design Turing criticized Jefferson's *a priori* assumption that, if the nervous system is continuous and digital computers are discrete, then the latter cannot imitate the former. We may now move to Descartes's assumptions, one of which is tackled in the imitation game itself since it has been designed from the start on the basis his language test.

Omitted among the passages that Jefferson quoted from Part V of Descartes's *Discourse*, there lies a central concept of Descartes which composes his famous, now near four hundred years old, beast-machine thesis. In a sentence that Jefferson did not include in his citation (the first sentence quoted below), Descartes attributed a sort of universality power to human *reason*:

> For whereas reason is a universal instrument which can be used in all kinds of situations, these organs need some particular disposition for each particular action. From which it comes that it is morally impossible that there be enough diversity in a machine for it to be able to act in all the occurrences of life in the same way that our reason would cause us to act. (DESCARTES, 1985, p. 140).

The first sentence was *not* part of the text that we know Turing to have read and marked. But the second one was, and it carries Descartes's corollary: having nothing like our reason, machines can't have enough diversity to be able to act in all the possible occurrences of life. Now, Turing was developing an argument that goes right on target against the view that machines can't have reason.[4] But before proceeding to Turing's view, let us complete the examination of Descartes's.

Descartes ended up associating reason or thinking with the special gift of a *rational soul*. It would be endowed by God only to the human species. At the same time, though, he considered that machines can't think and held this as a view that could be confirmed anytime by his language test. So, one may ask, was the metaphysical or the empirical perspective that would have made up his view? Descartes scholar John Cottingham seems to have posited it clearly:

> [T]he soul is "tacked on" at the end of the story [Cottingham's note: the *âme raisonable* is introduced right at the end, both in the order of exposition in the *Traité de l'homme* and in the summary recapitulation presented in Discourse Part V] invoked to account for the phenomena of thought and language that appeared to Descartes, for empirical reasons, radically resistant to mechanistic explanation. Whether that resistance can be overcome by the theoretically more sophisticated and empirically far richer resources of modern neurophysiology remains to be seen. (COTTINGHAM, 1992, p. 253)

So Cottingham implied that Descartes's notion of rational soul was rendered as an afterthought given the limits of his science, namely, the mechanistic "natural philosophy" developed in his posthumously published *The world* and *Treatise on man* (1985, p. 79). Descartes could not have foreseen electricity and the universality property of digital computers, indeed. He thus postulated a theoretical entity — the rational soul — as the causal element that could make an agent to be able to sustain a reasonable and unrestricted conversation.

Turing addressed Descartes's metaphysical assumption of the rational soul in his (1950) discussion of the first objection which he labeled "theological" (p. 443). Nonetheless, he did not do so by means of the imitation game. It turns out that, among the nine objections that Turing formulated and rebutted in his 1950 paper, the two first (the theological and the one he labeled "heads in the sand") are among those that he had already covered in (2004 [1948]; see Figure 1). Turing wrote that, "being purely emotional," these reactions to the possibility of machine intelligence "do not need to be refuted" (p. 410-1). In (1950) Turing restated this in the

[4] In Robert Stoothoff's English translation of the *Discours de la méthode* (DESCARTES, 1985, Part V, p. 131-41), the terms "thinking" and "thought" occur (p. 134), respectively, as translations for Descartes's use of French *penser* and *la pensée*; and the term "intelligence" occurs (p. 141) as translation for *esprit*. I consider that Descartes's notion of reason can be approximated — not unsoundly indeed — to thinking and intelligence.

end of his discussion of these objections (p. 444) and, overall indeed, he did *not* discuss them by means of the imitation game.

Descartes's corollary that machines lack enough diversity of behavior, however, Turing did criticize through the imitation game. In the third and four experimental questions (Q''', Q'''') that he asked about the possibility of a digital computer to perform well in the imitation game, Turing required the computer to be provided with a certain "storage capacity" and with "an appropriate programme" (1950, p. 442). Given these two properties, he declared to believe that the machine would be able to play the imitation game reasonably enough to deceive the interrogator in a rate of 3/10 after five minutes of interrogation. That is, it would be able to show sufficient diversity of behavior in the sense of the imitation game. Further on, in his discussion of objection five, the "arguments from various disabilities," Turing stated most explicitly:

> The criticism that a machine cannot have much diversity of behaviour is just a way of saying that it cannot have much storage capacity. Until fairly recently a storage capacity of even a thousand digits was very rare.
> (TURING, 1950, p. 449)

Turing's concession the limited storage capacity of machines up to then relates interestingly with Cottingham's comment above. For him, Descartes's note about the lack of behavior diversity in machines was motivated empirically, not metaphysically. In fact, Turing seems to have felt that he was fixing this kind of obsolete science of machine intelligence such as Descartes's which was based on Renaissance automata. Note, nonetheless, that Descartes's observation is not unrelated to the Lovelace-Hartree thesis. Descartes also suggested that the diversity of machine behavior was limited by the possibilities prearranged in their organs by their operator. And Turing would address all of this in his positive discussion. I shall now conclude my account of the Turing-Descartes connection in preparation for the heuristic function of the imitation game.

Turing had shown that digital computers hold the universality property (1950, §5, pp. 439-42), that is, such a computer can imitate the function of any organ or special-purpose machine by just being loaded with their program, with no need of redesign. In February 1947, in fact, Turing had referred to the capacity of the "universal machine" to "imitate" any other machine:

> Let us now return to the analogy of the theoretical computing machines with an infinite tape. It can be shown that *a single special machine* of that type can be made to do the work of all. It could in fact be made to work as a model of any other machine. The special machine may be called the universal machine; it works in the following quite simple manner. When we have decided *what machine we wish to imitate* we punch a description of it on the tape of the universal machine. (TURING, 2004 [1947], p. 383, emphasis added)

Turing seems to have got close to Descartes's notion of human reason as a universal instrument. And yet against the assumption of its uniqueness in nature, Turing had proven in (1936) that a single special machine could imitate any other machine (§A.2.1). But let us examine the nuances.

Turing's abstract universal machine had the disposition to be a universal instrument, but that disposition may not effectively be realized unless the universal machine is actually loaded with special-purpose programs to deal with (in Descartes's words) "all kinds of situations." Note that the universal Turing machine does not render universal behavior but rather sort of a potentially universal catalog of standalone particular behaviors. This would be similar to the lookup table that McCarthy and Shannon imagined, a sort of big mechanical parrot that, as they said, "does not reflect our usual intuitive concept of thinking." The logic of orchestration for the programs in the catalog, say, in prompt reaction to any one particular situation, would be still missing. And it was just that logic, I interpret, what Descartes called "reason." And Descartes associated reason with the ability to think. Turing's 1936 result did not yet address Descartes's assumption, and he seems to have understood this, for Turing never claimed his 1936 universal machine to be proof that machines can think. So why would have him proposed an imitation game that incorporated Descartes's language test as its cornerstone?

The short answer is that, for Turing, although his consolidated results would not yet be enough to establish that machines can think, they were suggestive enough that with some further decades of research machines would think. To further extend this answer, let us digress a bit. It turns out that Turing was working at least since late 1946 on the problem of how to make machines to learn like a brain. He described his plans in a *c.* November 1946 letter to cybernetician Ross Ashby (§A.3.4). Later in his 1948 NPL report (§A.3.9), Turing followed through with those plans. His 1948 text presented a view of *machine learning* — the first text ever on this topic — based on an analogy with the human brain in general and with the infant human cortex in particular. This was an articulation of a second computation model that he referred to as "unorganized machines." This one differed significantly from the 1936 model which was based on inserting special-purpose programs into the tape of the universal machine. While the universal machine proceeded by strict discipline to imitate each special-purpose machine that is loaded into its tape, an unorganized machine would proceed by initiative and learn from experience. It would be inspired on the learning of a child, and implement a connectionist computation model to learn how to behave both from being instructed explicitly by a teacher and from experience directly. In his 1950 text, by implicitly considering the "unorganized machine" computation model, Turing would offer a new discussion of his research agenda on the science of how to program "learning machines." Turing believed that this was the path towards making machines to think, now to definitively overcome Descartes's observation about the limited diversity of machine behavior and the Lovelace-Hartree thesis altogether.

The heuristic function

Recall that the heuristic function is to suggest and illustrate — for brevity, let us say *represent* — properties of the studied phenomenon in the thought experiment. Now, the imitation game is an experiment about the phenomenon of thinking or intelligence. So, if Turing designed it

to propose a new way to think of this phenomenon, what properties did him represent? For a short answer, I say that Turing represented two properties in his thought experiment. First, he suggested that the intelligence of an entity will always be a result of its *learning*. Second, he represented that the attribution of intelligence is a result of *wonder* in the observer at the time of the attribution. My proposition that Turing aimed in fact at the conceptual development of these two properties as constitutive of the phenomenon of thinking or intelligence have been object of a focused and detailed study before (§§2.3, 2.4). My goal here is to introduce them only in terms of their connection with the heuristic function of the imitation game. So a more specific question is how did Turing represent these properties within the imitation game? I find that each of the three schemes that compose the imitation game (§3.6), namely, (i) Descartes's language test, (ii) the blind communication arrangements and (iii) the presence of a human third player, were instrumental for the representation of those two properties. Now, before we go on and develop this further, let us remind of the core conceptual questions that were posed to Turing by his interlocutors. Some of them, he either dismissed or criticized (§3.7). But there were other questions that he rather saw as in effect in the open and in need of being positively addressed. These are the questions that he represented in the imitation game heuristically, as discussed next.

According to the Lovelace-Hartree thesis, a machine can do (only) whatever we know how to order it to perform. Hartree had argued in his *Calculating instruments and machines* (1949) that the use of the term "electronic brain" to refer to equipment such as electronic calculating machines "would suggest to the layman that equipment of this kind could 'think for itself,' whereas this is just what it cannot do" (p. 70). Hartree had thus expressed skepticism towards a property of digital computers that Turing considered to be feasible. Also, Turing was working with the game of chess to explore, illustrate and test for machine intelligence. But Polanyi was unimpressed by chess, and classed it as an art that "can be performed automatically" because its rules "can be specified" (1974 [1958]). Hartree had also suggested something similar. "All the thinking," he argued, "has to be done beforehand by the designer and by the operator who provides the operating instructions for *the particular problem*" (1949, p. 70, emphasis added). In passing, Jefferson had made the same point in his Lister Oration (1949a): "[i]t can be urged, and is cogent argument against the machine, that it can answer only problems given to it, and, furthermore, that the method it employs is one prearranged by its operator." He completed: "[t]he 'facilities' are provided and can be arranged in any order by 'programming' without rebuilding" (p. 1109). Chess was a particular problem, indeed. It would be difficult to overcome Polanyi's point and the Lovelace-Hartree thesis (brought by Jefferson as well), having chess as the target intellectual task to showcase machine intelligence. As of late 1949, Turing was cornered. How could one show that some particular behavior of a machine is truly a result of its own intelligence and not just a reproduction of the instructions thought out by its designer beforehand?

The answer, for Turing, I interpret, was in the composition of the three schemes that he built into his imitation game. Their combination allowed Turing to suggest and illustrate the two key properties that he associated with thinking or intelligence — learning and wonder. Turing

saw these properties themselves also combined, as if two sides of a coin. Let us see.

True machine intelligence, for Turing seemingly in concession to Hartree, would only be possible by reproducing the process of learning in analogy with human learning. Turing envisioned to submit machines to teaching and to learning from experience in general. If a machine is created with a very simple child program and can be taught in a similar way that a human child is to the extent that eventually it may take us and even its own teacher by surprise, then the Lovelace-Hartree thesis will be empirically shown to be false. And Turing made it clear in (1950) that this would be "[i]ntelligent behaviour" also because it "consists in a departure from the completely disciplined behaviour involved in computation" (p. 459). One would not be able to say anymore that machines can only do what they are instructed to do or that what they do is "purely mechanical behavior" (meaning mindless). Descartes's language test can in principle sense that, given the almost universal breadth of scope it implies. Since the interrogator is able to shift all of a sudden the conversation to *any* topic, there can be really no way in principle — except by the exotic idealization of a universal mechanical parrot — for the "operator" to "prearrange" the machine beforehand. The machine will have to think on the fly. So this view of thinking as a result of learning is represented in the imitation game. This process, Turing emphasized, is erratic. He thus consolidated his non-obvious view of intelligence: "[p]rocesses that are learnt do not produce a hundred per cent. certainty of result; if they did they could not be unlearnt" (*Ibid.*). Analogously to the performance of human intelligence, Turing held, it is by allowing machines to make mistakes that they will be able to show true intelligence.

Nevertheless, because machines are designed by ourselves, critics could try to downplay machine learning as genuine by pointing to some cause-and-effect explanation in terms of the machine's architecture and method. Turing advised about this with direct reference to Jefferson in the context of his discussion of objection "(5) Arguments from various disabilities:"

> The criticisms that we are considering here are often disguised forms of the argument from consciousness. Usually if one maintains that a machine can do one of these things, and describes the kind of method that the machine could use, one will not make much of an impression. It is thought that the method (whatever it may be, for it must be mechanical) is really rather base. Compare the parenthesis in Jefferson's statement quoted on p. [445-6, namely, "(and not merely artificially signal, an easy contrivance)"]. (TURING, 1950, p. 449-50)

For a "fair play," therefore, the decision of whether or not the machine exhibits intelligence of its own (beyond the intelligence of its designer as encoded in its program) should not be based on the machine architecture and method. This rationale would be just the same as in judging the intelligence of, say, a human child. The latter is not judged by studying their brain mechanisms. Nor is it reduced to the intelligence of a teacher or parents either. So true intelligence, for Turing, in addition to be a result on learning, was associated with the possibility of causing wonder in an observer who would then be surprised to receive a certain interesting response from an inquired entity. In 1948, Turing had written perhaps the most clear of his related notes about this:

> The view (e) that intelligence in machinery is merely a reflection of that of its creator is rather similar to the view that the credit for the discoveries of a pupil should be given to his teacher. In such a case the teacher would be pleased with the success of his methods of education, but would not claim the results themselves unless he had actually communicated them to his pupil. He would certainly have envisaged in very broad outline the sort of thing his pupil might be expected to do, but would not expect to foresee any sort of detail.
> (TURING, 2004 [1948], p. 411-2)

In his (1950) paper, then discussing "Lady Lovelace's objection" directly, Turing suggested one possible source for the view, perhaps more common among "philosophers and mathematicians," that "machines cannot give rise to surprises" (p. 451). It was "the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it." He even rephrased it as the assumption that "there is no virtue in the mere working out of consequences from data and general principles." Turing contended that this assumption is false. We could even say that he again associated himself, thereby, with an empiricist tradition.

Learning and wonder, for Turing, I interpret, were strongly coupled to one another. They were constitutive of what we call "thinking" or "intelligence," and composed, accordingly, Turing's theory of intelligence. Let us now see how a well played imitation game, whose structure is made of the combination of Descartes's language test, the blind communication arrangements and the presence of a human third player, may succeed at demonstrating learning by wonder.

Polanyi seems to have been right in that learning from experience in general — and this would be a central topic in the 1952 BBC roundtable (§A.4.6) — cannot be shown in arts whose rules can be specified. Hartree and Jefferson's references to "instructions for the particular problem" and "only problems given to it" also suggested that a general problem was needed to test for machine intelligence. It had to be an intellectual task that is representative of human intelligence and could not be discredited as such. And this is just what Descartes had already established over three centuries before, namely, that for machine intelligence to match human intelligence it would have to be demonstrated within an almost universal breadth of scope, preferably through a language test. Turing seems to have noticed. Besides, Descartes's language test was brought forward and promoted by Jefferson himself so it could not be challenged by him. The cartesian approach, Turing seems to have conceded, was sensible for testing whether a machine is learned and can learn from experience in general. For one key feature, a unrestricted conversation allows for sudden changes of intellectual field. Turing illustrated this by showing a few specimen questions and answers about sonnets, arithmetics and chess (§3.6).

And yet Descartes's language test, as it was, would not offer a fair test. Observers such as Hartree, Polanyi and Jefferson, who may have been decided *a priori* about the possibility of machine intelligence, could fall prey to confirmation bias. So a first reshaping that Turing made was to make it a blind language test. The arrangements for blind communication addressed bias about seeing, touching and hearing the contestant entity. By these means critics could not have access either to the machine's method, thus made a black box. In case it performed well,

one would not be able to deny its intelligence just because there may be a cause-and-effect explanation of its method, "whatever it may be," Turing argued, "for it must be mechanical." But here comes a caveat. If the machine's architecture and method could not be inspected, what if it was made not of any learning but mere trickery? Turing suggested that the machine should try to deceive the interrogator because, of course, its goal in the game is to pretend to be of a species and gender that it isn't. He did not suggest, however, the designer to cheat on the imitation game which was meant to test for true machine intelligence. About this, in his 1950 text Turing wrote:

> What would Professor Jefferson say if the sonnet-writing machine was able to answer like this in the *viva voce*? [...] [I]f the answers were as satisfactory and sustained as in the above passage I do not think he would describe it as 'an easy contrivance'. This phrase is, I think, intended to cover such devices as the inclusion in the machine of a record of someone reading a sonnet, with appropriate switching to turn it on from time to time. (TURING, 1950, p. 446-7)

Later in his *c.* 1951 BBC radio lecture, Turing emphasized that the machine learning processes that he envisioned "could probably be hastened by a suitable selection of the experiences to which [the machine] was subjected" (2004 [*c.* 1951], p. 473). "But here we have to be careful," Turing pondered. He alerted about the "easy" approach of arranging the experiences of the machine "in such a way that they automatically caused the structure of the machine to build up into a previously intended form." And this, Turing reprehended, "would obviously be a gross form of cheating, almost on a par with having a man inside the machine." In light of this, the criticism of Turing's test on account of its vulnerability to the ELIZA effects (Cf. Weinzenbaum 1966) — which come out when the machine is programmed not to learn from experience but rather to apply psychological trickery — does not seem sensible. We may consider that the imitation game is a test for true machine intelligence and it has not been designed to be immune to scams, just like experiments in physics are not designed to prevent trickery from the part of the scientists in charge.

And yet one may insist that, whatever learned behavior a seriously-programmed machine may exhibit, it must be still just a result of the intelligence of the programmer. Turing's response, in that case a half satiric one, is a shift of the burden of proof towards the human side for balance. This is represented in the imitation game by the introduction of a human third player that is also a contestant in the game, a woman or a man depending on which of its versions is considered. The human contestant can contribute to debunk the machine on the fly during the game. If the machine is just reproducing a pre-programmed behavior, how likely will it be able to adapt to the challenges posed to it by both the human player and the interrogator? How likely will it be able to cause wonder in connection to situations that arise within the conversation? In any case, if the machine contestant has to prove the unprovable, the same will hold for the human contestant beside it. This is because the interrogator will not be able to know *a priori* who is who.

Altogether, Turing proposed the imitation game as a means to ground the discussion on the possibility of machine intelligence. Essentially, I have interpreted, Turing relied on his

thought experiment to represent learning and wonder as key properties of the phenomenon of thinking or intelligence. I would like to counterpoint two aspects of Hayes and Ford's (1995) criticism in particular. They dismissed the goal of Turing's test by claiming that it discriminates non-human (say, other animal) forms of intelligence. And they saw in the test no value "as a guide to research" either, claiming that it provides no way to measure partial progress toward the goal (p. 973). The test has often been accused, in fact, of trying to establish human intelligence as the one standard of intelligence at the expense of other forms of intelligence in nature. But this claim is ahistorical too. If there were species chauvinists in the British public debate about the new electronic computing machines in the late 1940's and early 1950's, Turing, as we have seen, was not one of them. Quite the opposite. Intelligence, in Turing's view as we have seen, is constituted by the property of learning. And from the lower to the higher forms of intelligence in nature, we know that learning — by reflexes or whatever nervous-system mechanism — is its common and overarching basis. Besides, in connection with the claim that the test has no use in guiding research, there may be another aspect of Ernst Mach's analysis of thought experiments (1976 [1897]) that is worth of attention here. Mach argued that "thought experiment often precedes and prepares physical experiments" (p. 136). From this point of view, Turing's imitation game may be seen by scientists in terms of what its heuristic function has taught about the studied phenomenon. I claim that it could have guided and still can guide both (i) research itself by taking into account his view that intelligence is a result of learning; and (ii) the design of physical experiments to evaluate research progress by taking into account his view that the attribution of intelligence is a result of wonder.

Concerning learning from experience in general, Turing's goal was to make a machine to pass a variant of Descartes's language test. I think that the research program that John McCarthy built to address the representation of commonsense knowledge partially intersects with Turing's.

Concerning wonder, there has been a historical tendency from the part of AI scientists to dismiss it (then understood as crude deception) as an irrelevant part of the imitation game. I think that Diane Proudfoot argued this neatly in (2011). Scientists may not yet be familiar with the conceptual problem Turing was addressing. This may be also because they did not yet effectively face it, as their systems are not yet mature to be tested for true machine intelligence. From a scientific research point of view, there would be no problem in opening the box of a machine contestant to study whether it is just a mechanical parrot. But as Turing argued (and we have seen his argument in depth in §2.3), once one knows the mechanism underlying the performance of an entity, they tend to fall prey to bias towards not attributing intelligence to it anymore. In fact, much time later in (1985), Marvin Minsky presented a definition of intelligence that matched the wonder component of Turing's view of intelligence. For him,"intelligence" is "our name" for whichever mind processes that "solve problems we consider difficult" and "don't yet understand" (p. 71). He thus added:

> Some people dislike this "definition" because its meaning is doomed to keep

> changing as we learn more about psychology. But in my view that's exactly how it ought to be, because the very concept of intelligence is like a stage magician's trick. Like the concept of "the unexplored regions of Africa," it disappears as soon as we discover it. (MINSKY, 1985, p. 71)

Minsky allusion to Africa reflects a colonialist view of the world which is in clear contrast with Turing's two observations quoted at the end of my study of his irreverence (§1.2). In spite of that, in any case, Minsky seems to have caught Turing's view that the designation of intelligence is an elusive phenomenon. And this is in fact a precise conceptual response to one of Hayes and Ford's complaints about Turing's test. They wrote:

> The species [*machine-imitates-man*] test further reveals the poor experimental design of the imitation game in the difficulty of obtaining an unbiased judge. The general perception of what are essentially human talents keeps shifting. As AI progresses and more and more tasks previously considered to involve human abilities are performed by machines, a judge in the naive Turing Test will gain more and more subtle ways of detecting the behavior of nonhuman machines, just as a skilled doctor will become more adept at recognizing subtle symptoms. (HAYES; FORD, 1995, p. 974)

In my view, Hays and Ford's criticism about the bad experiment design of the imitation game is fair only if we see it as the proposal of an actual experiment. Their criticism loses strength as we see Turing's proposal as a thought experiment whose heuristic function is to represent the properties of learning and wonder. In light of the heuristic function of Turing's test, perhaps, Hayes and Ford could have appreciated its scientific and philosophical value.

In sum, I hope to have succeeded in my construction of Turing's imitation game as a thought experiment. We have studied its source materials and inner structure, and its critical and heuristic functions. All elements were set to convey, once neatly combined in the full-fledged thought experiment, Turing's scientific and philosophical argument within the mind-machine controversy. In light of all that, I shall now further summarize this proposed view of the imitation game as a thought experiment to address the Turing test dilemma.

3.8 Turing's imitation game is a thought experiment

We shall now be ready to revisit the question of the Turing test dilemma: did Turing propose his test as an experiment to empirically decide for his question, "can machines think?"

As anticipated, my answer is reconciliatory: yes and no, in the sense that Turing, in my view, proposed his test as a *thought experiment*, no matter he was conscious about that or not.

I shall now argue for the two horns of this dilemma to drop out. Let us recall them. By the first horn, let the test be taken to be an experiment. Then how could we explain that it has been tried in practice as such and has been generally argued to be either underspecified or just a piece of bad experiment design? If the imitation game is seen as an experiment to "replace"

the original question, what exactly is the new (experimental) question that Turing expected the imitation game to answer? We shall now look at this closely. Contrariwise, by the second horn, let the test be dismissed as an experiment. So this would mean that Turing's proposal boils down to a piece of rhetorics with no scientific meat inside. And yet Turing does seem to have proposed the imitation game (test, or experiment) to "replace" the original question, which he suggested that was "too meaningless to deserve discussion."

I shall refer first to the interpretive basis I have offered for Turing's 1950 text (§3.3), and to the epistemological (§3.4) and historical (§3.5) roots of his proposal, and then proceed to address the two horns of the dilemma as a corollary of my construction of the imitation game as a thought experiment (§§3.6, 3.7).

We have seen that Turing's 1950 proposal was presented from within a scientific controversy on mind and machine, notably with Douglas Hartree, Michael Polanyi and Geoffrey Jefferson. Turing's direct and indirect discussion with these three thinkers is key for any exegesis of Turing's 1950 paper in general, and to an understanding of the conceptual problems he tried to solve in his "imitation tests" in particular. Until October 1949 Turing had worked with the game of chess as intellectual task to illustrate, explore and test for machine intelligence in terms of a physical (actual) experiment. This was a choice for convenience over appeal. Also light of epistemological constraints on his existential hypothesis, he needed to showcase an actual thinking machine. However, pressed to respond to the non-trivial challenges posed by them, Turing made his crucial 1949 move. He sacrificed his preference for convenience and empirical feasibility and opted for appeal instead. He traded the possibility of showing machine intelligence in a physical experiment based on chess-playing in his own lifetime for arguing about it in a thought experiment based on conversational question-answering which he knew very well that could only be convertible to physical experiments in the future.

We may now address the difficulties of facing the two horns of the Turing test dilemma.

I shall start with the second horn. In the exegesis of his 1950 text, we have seen that Turing formulated his question — "can machines think?" — as an empirical problem, and discussed it by means of the imitation game. Turing's reference to "criterion for thinking" is a sign that his discussion was meant to have some epistemological significance. I presented evidence that Turing's "discussion" was no casual initiative of his and no vacuous rhetorics. Rather, as a reader of Bertrand Russell, he seems to have felt compelled to follow a well-known method in the history of philosophy, namely, the question and answer or Socratic dialectic method. He employed this method with a clear philosophical goal in mind, which was to fix what he saw as logical errors of his opponents, notably Hartree, Polanyi and Jefferson. But he respected the empiricist boundaries indicated by Russell to have been exemplified by Galileo. As he conducted such discussion by means of the imitation game, he varied its design. So he did not specify any one particular design for a definite and practical experiment indeed. And there is no reason to suppose that his sketchy narrative was a result of inadvertent imprecisions. Rather,

as we have seen, he let the design of his imitation-game experiment to deliberately slip through various versions from 1948 to 1952.

Let us now shift to the first horn. The charges of underspecification and bad experiment design are unjustified if it is noticed in Turing sources (1948-1952) that there is no such thing as one particular version of Turing's imitation game or test that is preferable over the others. In agreement with what we have seen in connection with the second horn, the several variants of question that Turing asked discursively were meant for offering a general empirical basis for discussing the original question under different conditions. However, that does not mean that Turing's proposal contained no idea of experiment to test for machine intelligence at all. The point is rather that his experiment was (deliberately) flexible by design and tailored to address conceptual problems. Turing varied its specification continuously through discourse to stress key properties of the phenomenon under different conditions. Now, in the history of science, this continuous variation of experimental conditions is no vacuous rhetorics. According to one of the scholars to have firstly used the term "thought experiment," Ernst Mach, this is rather a key property of thought experiments:

> The outcome of a thought experiment, and the surmise that we mentally link with *the varied conditions* can be so definite and decisive that the author rightly or wrongly feels able to dispense with any further tests by physical experiment. However, the less certain their outcome, the more strongly thought experiments urge the enquirer to physical experiment as a natural sequel that has to complete and to determine the result. (MACH, 1976 [1897], p. 137-8, emphasis added)

Mach suggested, I interpret, that when the conditions of a thought experiment are varied enough for the study of a certain conjecture, the outcome may be so definite and decisive that one may feel endowed to dismiss any further physical experiment. Turing, as we have seen, did not dispense with physical experiment. He abided by Russell's empiricist guidelines. In any case, Mach considered paradigmatic examples. Let us then gain more depth into his point:

> Conditions that have been recognized as of no account with regard to a certain result can be varied at will in thought without altering that result. By astute handling of this procedure we may reach cases that at first blush seem rather different, that is to generalisation of the point of view. Stevin and Galileo showed great mastery of this device in their treatment of the inclined plane. Poinsot, too, used this method in mechanics. To a force system A he adds two others, B and C, C being chosen to balance each of A and B. Since the observer's point of view is irrelevant we are led to recognize A and B as equivalent, although they might differ greatly in other ways. Huygens' discoveries about impact rest on thought experiments: starting from the knowledge that the motion of other bodies is as irrelevant to the colliding body as it is to the observer, he changes the observer's point of view and the relative motion of the surroundings: in this way he starts from the simplest special case and reaches important generalizations.
> (MACH, 1976 [1897], p. 138)

Turing's discussion, I claim, follows precisely the approach described by Mach. He came up with a first experimental scenario (the *man-imitates-woman* game) to which he invited us to think

about, and kept varying this scenario and asking questions that ranged from Q' up to Q'''', as we have seen. I have pointed out key elements of Turing's various experimental scenarios and their functions (§§3.6, 3.7). Mach thus summarized his characterization of thought experiments:

> It is further profitable mentally to vary those conditions that are decisive for the result, the most fruitful approach being continuous variation, which yields a conspectus of all possible cases. Thought experiments of this kind undoubtedly have led to enormous changes in our thinking and to an opening up of most important new paths of enquiry. [...]
>
> As we see, the basic method of thought experiments, as with physical experiments, is that of variation. By varying the conditions (continuously if possible), the scope of ideas (expectations) tied to them is extended: by modifying and specializing the conditions we modify and specialize the ideas, making them more determinate, and the two processes alternate. (MACH, 1976 [1897], p. 139)

This is, I hold, in striking agreement with Turing's approach, namely, the "continuous variation" of the "conditions that are decisive for a result." Turing, as I suggest overall, used to run his imitation game experiment (from 1948 to 1952) according to various settings — in his mind.

I shall now proceed to complete this study by suggesting that Turing's imitation game is a thought experiment in science.

3.9 Turing's thought experiment vis-à-vis Galileo's

There is an analogy that I will suggest between Turing's imitation-game and Galileo's falling-bodies thought experiments. The analogy may be non-obvious, if for nothing else, because of two common views about Galileo's experiment which lie in the way of it and thus need to be dismantled before I can argue my point. These views have both long been (in spite of their different timescales) spread in philosophy and physics books and only recently have been undermined by rigorous scholarly work. I will start with the philosophy of science story, and then shift to the history of science one.

There is a view, stated for example by Popper (2002 [1959], p. 465) and recently repeated by Dennett (2013, p. 21) that Galileo's thought experiment carries a *reductio ad absurdum* argument showing Aristotle's theory of motion to be actually inconsistent. If this is so then there might seem to be little point in comparing the function of Galileo's thought experiment in the polemic with the peripatetics with Turing's imitation game, for the latter does not seem to perform a similar function in the mind-machine controversy. Essentially, Turing rebutted Hartree, Polanyi and Jefferson's assumptions not by showing that they were false in themselves but by exposing that they were *a priori* and that they could apply to the human likewise to machines. Turing's argument consisted partly in a shift of the burden of proof (§1.6), but also depended on the confirmation of his existential hypothesis on thinking machines (§3.7).

Besides, there is also a view that Galileo would have actually decided his contention by running his thought experiment, say, from the top of the Leaning Tower of Pisa. He would have

empirically (and publicly) demonstrated that his theory of motion was right and Aristotle's was wrong. If Galileo actually ran his experiment and thus fulfilled its epistemic function empirically, then this would be an alternative reason why there might make no sense to compare it with Turing's. For the latter, as we have seen, was not feasible to be actually run in Turing's lifetime (§3.4) nor were very effective either the actual attempts at running it later (§3.1).

To argue against both views of Galileo's experiment, I shall collect and combine some of the latest related scholarly works. After that, I will discuss science and engineering developments in connection with the feasibility of conditions that are idealized in Galileo's thought experiment. By analogy, this shall shed light on the feasibility conditions that are idealized in Turing's. I will conclude by comparing Turing and Galileo's thought experiments from a point of view of the philosophy of science.

Did Galileo's thought experiment draw Aristotle's theory onto a *reductio*?

Did Galileo refute Aristotle's theory of motion by logic alone? If so, it would mean that he settled the contention by force of the design of his thought experiment itself (or the argument underlying it) such that an actual or physical experiment would be needed no more. In the meantime in between Popper and Dennett, this view has been endorsed by Brown (1991, p. 1-2) Before going further, let us see the famous Galilean passage that would have granted the alleged refutation (it is important to keep track of the material, size, weight and speed of the falling objects):

> SALVIATI: But without experiences, by a short and conclusive demonstration, we can prove clearly that it is not true that a heavier moveable is moved more swiftly than another, less heavy, these being of the same material, and in a word, those of which Aristotle speaks. Tell me, Simplicio, whether you assume that for every heavy falling body there is a speed determined by nature such that this cannot be increased or diminished except by using force or opposing some impediment to it [...].
>
> [SIMPLICIO acquiesces]
>
> Then if we had two moveables whose natural speeds were unequal, it is evident that were we to connect the slower to the faster, the latter would be partly retarded by the slower, and this would be partly speeded up by the faster [...].
>
> [SIMPLICIO agrees again]
>
> But if this is so, and if it is also true that a large stone is moved with eight degrees of speed, for example, and a smaller one with four [degrees], then joining both together, their composite will be moved with a speed less than eight degrees. But the two stones joined together make a larger stone than the first one which was moved with eight degrees of speed; therefore this greater stone is moved less swiftly than the lesser one. But this is contrary to your assumption. So you see how, from the supposition that the heavier body is moved more swiftly than the less heavy, I conclude that the heavier move less swiftly. (GALILEI, 1974 [1638], p. 66-67)

Logical reconstructions of Galileo's argument in this passage have been presented by various commentators, notably Brown himself (1986), John Norton (1996), Tamar Gendler (1998), and Richard Arthur (1999). Their discussion is focused on what Norton cast as "the epistemological

problem of thought experiments in the sciences," namely, if thought experiments are supposed to give us information about our physical world, then from where can this information come? (1996, p. 333). This debate initiated by Norton's strong criticism of Brown's Platonist answer to the question. Galileo's falling-bodies thought experiment as contained in the above passage eventually became their paradigmatic case study. I am in no position here to take part in their discussion. Suffice to say that I do not accept Galileo's argument as a straightforward logical refutation of (in Popper's words) "the Aristotelian supposition that the natural velocity of a heavier body is greater than that of a lighter body." Norton claimed that thought experiments are nothing but picturesque arguments. For him, their epistemic value reduces to the soundness of the logical argument such that the thought experiment design and any non-propositional elements composing its scheme can be dispensed. While I do not side with this view either, in order to argue my point I shall borrow Norton's reconstruction of Galileo's argument. It runs as follows (I bracket the logical steps for emphasis):

> (1). Assumption for *reductio* proof: The speed of fall of bodies in a given medium is proportionate to their weights.
> (2). From (1): If a large stone falls with 8 degrees of speed, a smaller stone half its weight will fall with 4 degrees of speed.
> (3). Assumption: If a slower falling stone is connected to a faster falling stone, the slower will retard the faster and the faster speed the slower.
> (4). From (3): If the two stones of (2) are connected, their composite will fall slower than 8 degrees of speed.
> (5). Assumption: the composite of the two weights has greater weight than the larger.
> (6). From (1) and (5): The composite will fall faster than 8 degrees.
> (7). Conclusions (4) and (6) contradict.
> (8). Therefore, we must reject Assumption (1).
> (9). Therefore, all stones fall alike. (NORTON, 1996, p. 341-2)

Now, the final step (from 8 to 9), Norton pointed, "is actually quite tricky." (I shall now resume with Norton's elaboration and build upon it in the sequel.) There is a key tacit assumption behind it, which Norton thus formulated:

> "(8a). Assumption: The speed of fall of bodies depends only on their weights" (p. 342).

Norton then emphasized how ahistorical a claim is to purport that Galileo's *reductio* refuted Aristotle's view:

> This reading is, however, a serious misreading historically of what is actually at issue in the relevant part of Galileo's *Two New Sciences*. For Salviati is in no position to make assumption 8a. The context is the fall of bodies *in media* and Aristotle's view is clearly stated by Simplicio as applying to exactly this case. Modern readers, of course, are usually unable to resist the temptation of dismissing the medium through which these bodies are falling as a confounding distraction that should be idealized away. Are we not really talking about bodies falling in a vacuum and is this not what Galileo's theory is really all about? Perhaps, but at this point in the dialogue, Salviati is in no position to assume away the medium. The broader focus of discussion, the very point that raised

> the question of falling bodies, is the possibility of the existence of a vacuum. To assume the possibility of a vacuum at this point would be to beg the main question under discussion. (NORTON, 1996, p. 344, no emphasis added)

Whatever criticism Norton's reconstruction may receive, his point about the focus of Galileo's argument and of the discussion itself is indisputable. Indeed, Galileo's goal was to undermine Aristotle's alleged proof of the impossibility of motion in a vacuum after Simplicio had said: "Aristotle shows that it is precisely the phenomenon of motion, as we shall see, which renders untenable the idea of a vacuum" (1974 [1638], p. 61). It turns out that hidden assumption (8a) is empirically false *in media*.

Let us now go beyond Norton's reconstruction and observe Salviati's key move. "Tell me, Simplicio," said Salviati, "whether you assume that for every heavy falling body there is a speed determined by nature such that this cannot be increased or diminished except by using force or opposing some impediment to it" (1974 [1638], p. 66, emphasis added). Now, the reader may recognize what is in this passage? It is nothing less than *the principle of inertia or inertial motion*. Thereby Galileo had established the basis for his acclaimed *reductio*, and actually, one of the most fundamental laws of modern physics. It was Galileo himself who idealized the removal of the medium, whereas the presence of a medium was a premise of Aristotle's theory. Galileo's *reductio* is sound. While Norton's reconstruction in turn, by starting from assumption (1), does require assumption (8a) to be sound. Galileo's proof is *not* a refutation of Aristotle's assumption, as Popper and others claimed. Galileo himself did not. Indeed, as a peripatetic Simplicio could not have given assent to proposition Salviati's first move. Aristotle's theory prohibited motion in vacua, and this is exactly what Galileo sought to undermine.

We may now move to the history of science legend about Galileo's thought experiment.

Did Galileo ever run his experiment to empirically decide the contention?

The story is one of the most famous anecdotes in the history of science. Sometime around the year 1590, Galileo would have climbed the Leaning Tower of Pisa and dropped from the top two objects of different weights in order to disprove Aristotle's law of fall, which claimed that the speed of fall of bodies is proportional to their weight. By letting the two objects fall simultaneously and showing that they reached the ground simultaneously, Galileo would have thus demonstrated to the professors and students gathered around the tower that Aristotle was wrong. Now, is this story true or legend? I shall take a moment to inquire it, and the importance of the answer for the analogy with Turing's case shall become apparent at the end.

Various studies have addressed this problem, specially recent ones from the late 1970's to the late 1980's. (I say "recent" by considering the timescale of a story that is known to have been originated in 1654, only some 12 years after Galileo's death.) The best study I know of it is Michael Segre's (1989) well-balanced and fascinating historiography. I will mostly reproduce its findings and related discussion, but I shall give more attention to the discussion

with Stillman Drake and conclude by referring to the empirical results of studies by physics education researchers.

Unlike, say, the story of Newton's apple, the story of the leaning tower experiment has never been mentioned in Galileo's writings nor is there evidence that he ever narrated it. Galileo scholar Stillman Drake professed in (1978):

> [H]is demonstration from the Leaning Tower with bodies of the same material more than fifty years earlier was quite forgotten. Indeed, most historians believe that there was nothing to be remembered and that no such test as described in Viviani's biography of Galileo was ever made. I am now inclined to believe that if Galileo's answer to the above letter had been preserved, it would long ago have provided definite evidence to the contrary. (DRAKE, 1978, p. 415)

The letter mentioned by Drake is an inferred reply from Galileo to Vincenzio Renieri's 13 March 1641 letter. The reply has never been reported and yet is supposed to have contained the description of the leaning tower experiment. In Renieri's letter — cf. Antonio Favaro's (1968, p. 305-6) edition of *Le opere di Galileo Galilei* — two experiments are reported to have been conducted by Renieri at the top of a (some) tower. In the first experiment, a lead ball and a wooden ball (*n.b.*, the first is heavier) were dropped from the tower, the former reaching the ground first by at least three ells (about 1.75 metres). In the second experiment two lead balls were used, one the size of a cannonball and the other of a musket bullet (*n.b.*, the first is heavier again but now has the same density), with the larger reaching the ground first by a palm (about 22 cm). So if 22 cm is considered a significant distance, then both results are unfavorable to the conclusion — say, as formulated by proposition (9) in Norton's reconstruction of Galileo's thought experiment — that bodies of different weights would fall alike. In a second letter of Renieri's only seven days later (20 March 1641, cf. Favaro's 1968 collection, p. 310), he wrote: "but that two heavy bodies, unequal in weight but of the same material, falling from the same height perpendicularly have to arrive with *different* velocity and in different time at the centre, this I think I have heard or read from you" (emphasis added). From this, it is generally concluded that Galileo must have replied indeed. But Drake goes further and infers also the description of the leaning tower experiment from it. I shall come back to the discussion with Drake soon.

The story about the Leaning Tower of Pisa was actually reported by one of Galileo's closest assistants, Vincenzio Viviani (1622-1703), in his biography of Galileo published only posthumously in 1717. No other source has ever appeared. I reproduce Viviani's description from Antonio Favaro's (1968) edition of *Le opere* as selected and quoted by Drake in (1978):

> [A]nd then, to the great discomfort of all the philosophers, through experiences and sound demonstrations and arguments, a great many conclusions of Aristotle himself on the subject of motion were shown by him to be false which up to that time had been held as most clear and indubitable, as (among others) that speeds of unequal weights of the same material, moving through the same medium, did not at all preserve the ratio of their heavinesses assigned to them by Aristotle, but rather, these all moved with equal speeds, he showing this by

> repeated experiments [*esperienze*] made from the height of the Leaning Tower of Pisa in the presence of other professors and all the students.
> (VIVIANI, 1968, vol. 19, p. 606), (DRAKE, 1978, p. 19-20)

This was all that Viviani brought forth about the tower experiment, and there is no one source other than this ever mentioning an experiment from the Leaning Tower of Pisa. Segre studied (1989) Viviani's report in depth by considering his scientific, social and cultural context, and also reviewing the discussion in the secondary literature chronologically to arrive at the known truth about the story. I shall now take a shortcut to his conclusions.

Viviani's report has been accepted silently until the nineteenth century, when Galileo scholars Raffaello Caverni (1837-1900), a Florentine priest, and Emil Wohlwill (1835-1912), a German historian of science, firstly challenged it. Antonio Favaro (1847-1922), the notorious editor of *Le opere*, denied their (independent) criticisms on the basis of Viviani's supposed reliability as a source. This is essentially the same line of defense that would be made by Drake later. In the twentieth century a new wave of criticism has been initiated by Lane Cooper (1935). Soon after Alexandre Koyré's much deeper (1977 [1939]) *Galilean studies* came out. As known, Koyré suggested that experiment did not play any essential role in Galileo's science and that some experiments described by him may have never been actually done. (I only found that Segre has been mistaken about his thought, p. 443, that Koyré did not question the truth of the leaning tower anecdote in particular, for this is just what he did in his lesser-known 1937 "Galilée et l'expérience de Pise.") Galileo drafted his *De motu* from 1589 to 1592 when he was still in Pisa. In these writings he discussed tower experiments many times and even a preliminary version of the famous thought experiment that would appear later in *Two new sciences* (1974 [1638]), but he did not outline any precise experimental description. In fact, he repeatedly stated that bodies of different weights fall with *different* speeds, and even went to write: "[t]his is something I have often tested" (1960 [1590], p. 107). In the lack of sources other than Viviani's, Segre found a most fruitful approach by shifting the problem from "did Galileo perform the leaning tower experiment?" to "Why did Viviani think it important to report such an experiment?" (p. 444). By inquiring into Viviani's context, Segre found in his *Life of Galileo* a strong influence of art historian Giorgio Vasari (1511-1574) whose standard was to embellish artists' images by means of anecdotes. Segre also found cultural clues about Viviani's readership (p. 446-7) and material evidence on Viviani's embellishment approach (p. 447-8). "Most of what Viviani says," Segre remarked, "is to a large extent true, with a small degree of fiction — an embellishment imposed on him by the literary conventions of his day" (p. 449). We do have evidence that Galileo climbed a famous tower to make a demonstration. He showcased his telescope to authorities at the top of the Tower of San Marco in Venice. Also, other professors at Pisa did make experiments on falling bodies from the leaning tower. For instance, Giorgio Coresio would have done so in 1612 and claimed that the bodies behaved exactly according to Aristotle's law. His goal was to disprove Jacopo Mazzoni's 1597 opposing claims, which he said to have been based on experiments performed from Mazzoni's window, which he claimed in turn not to be high enough

(p. 450). Segre acknowledged the value of the legend for the sake of methodological discussions in historiography of science, but concluded that the leaning tower experiment "should not be a matter for consideration by historians of science, but rather by historians of literature" (p. 451).

Drake would not agree. He kept a longstanding and well-dressed view of Galileo's use of experiment, and claimed that the accumulated criticisms concerning it were largely unfounded. For a critical discussion of Drake's position, the reader may refer to Naylor (1974). He showed that Galileo's attitude to observation may well have been far more complex than Drake supposed. In the development of his historical and scientific analysis, Naylor built upon James MacLachlan's remark that continuing disagreement over Galileo's use of experiment should lead to further (empirical) examination of Galileo's experimental claims. Now, this is exactly what we shall do at last, but not least, again by referring to the secondary literature.

So, whatever is the truth of the legend, what if Galileo had actually done the leaning tower experiment? Could he have obtained the results claimed? If so, this would mean that Galileo's thought experiment would have been functional as an actual experiment to decide for the contention in his own lifetime. It turns out that Carl Adler and Byron Coulter asked just that question (1978). They presented a concise and yet rigorous study on both the history and the physics of the problem. They quoted various Galileo's passages from *De motu* and *Two new sciences* as we have seen, and discussed possible interpretations of the experimental settings implied. Specifically, they quoted from Galileo (1974 [1638]): "[t]his means that a lead ball falling from a tower two hundred braccia high [about 100 meters] will be found to anticipate an ebony ball by less than four inches [approximately 0.102 meters]" (p. 79). This is one of the most clearly specified experimental settings given by Galileo in the *Two new sciences*, which is his most mature work published in 1638. According to Adler and Coulter, in other passages of the *Two new sciences* Galileo suggested that he knew that in air a heavy ball would fall faster than a light ball of the same size, but he underestimated the effect of air resistance. They set out to reproduce it in the laboratory and even recorded the experiment. For one setting, they dropped an iron ball and a rubber ball of same size from a height of 38 m. They found that "[t]he lighter ball is seen to be more than 7 m high when the heavy ball reaches ground" (p. 200). The motivation to use these specific settings was to compare with Gerald Feinberg's predictions for this phenomenon (1965), because Feinberg also gave calculations considering the height assumed by Galileo (100 m) which is harder to reproduce. Their physical results matched Feinberg's prediction for the 38-meters high run. Now, Feinberg's prediction considering instead a lead ball and an ebony ball each with a radius of 10 cm being dropped from a height of 100 m — which matches exactly Galileo's setting — was that the ebony ball would be more than 5 m high when the lead ball hit ground. "This is far more," Adler and Coulter observed, "than the 4 in. [approx. 0.102 m] claimed by Galileo in the earlier quotation." These results show, in sum, that Galileo could not have obtained proposition (9) as stated in Norton's reconstruction from the Leaning Tower of Pisa. Adler and Coulter further commented:

> If Galileo had used two balls of the same material, but of different sizes and weights, as we have shown he probably would have done in the legendary tower experiment, the balls would have hit more nearly simultaneously than in the [above example]. The effect of the greater weight of the larger ball would have been partially balanced by the effect of its greater size. Nonetheless, two balls falling through the air would actually hit at the same time only if the product of the density and radius of each ball were the same. Although we do not know what size balls Galileo supposedly used in his experiment, we do know that in *Two new Sciences* he claimed that if a 100-lb. iron ball and a 1-lb. iron ball were dropped from a height of 100 braccia, then "[...] when the larger one strikes the ground, the other is two inches behind it." (GALILEI, 1974 [1638], p. 68). (ADLER; COULTER, 1978, p. 200)

Finally, Adler and Coulter cited physics textbooks and their dissemination of the story. The problem, they argued, is not that physics teachers refer to the legend of the leaning tower. It is rather that they often suggest that Galileo would have obtained proposition (9). But this is false.

There have been, of course, other studies coming to the same conclusions. For a most comprehensive and detailed considering even the particular setting of tying the heavier falling object to the lighter one as described in the allegedly *reductio*, see Christensen et al. (2014).

We shall now be ready to suggest the analogy between Galileo and Turing's thought experiments.

The dual function of Galileo's thought experiment

Overall, Brown offered this interpretation about the dual function of Galileo's falling-bodies thought experiment: "Galileo's account of free fall did two distinct things: first, it destroyed Aristotle's view that heavier objects fall faster; and second, it established a new account that all objects fall at the same speed" (1991, p. 43). He then added this side note: "(Thought experiments are fallible, of course, so my use of terms 'destroy' and 'establish' should be understood as merely tentative.)" In fact, Galileo's falling-bodies thought experiment did *not* refute Aristotle's law of fall and yet it performed its critical function of questioning it. It can hardly be said to have established a new account that all objects fall at the same speed unless, as Brown seems to concede, we take "establish" with a grain a salt indeed. Now, what else could we say about the heuristic function of Galileo's thought experiment?

I am certainly in no position here to engage in a full-fledged discussion of the research program promoted by Galileo in the *Two new sciences*, which can be found in the studies of Galileo scholars and of historians of early modern science in general. I shall comment it only in connection with the reconstruction of Galileo's argument as conveyed in his thought experiment. In my view Norton (1996) caught the central point of the heuristic function of the experiment here: "[t]he broader focus of discussion, the very point that raised the question of falling bodies, is the possibility of the existence of a vacuum" (p. 344). That is, *Galileo's contention, just like Turing's, was to be allowed to pursue an existential hypothesis*. The possibility of motion in vacua required, first, the existence of vacuum, and second, the confirmation of a key property

of it, namely, to afford motion such that bodies of different weights move through it with equal natural speed. In fact, in a further passage of *Two new sciences*, Salviati said:

> This seen, I say, I come to the opinion that if one were to remove entirely the resistance of the medium, all materials would descend with equal speed. (GALILEI, 1974 [1638], p. 75)

In short, Galileo presented his falling-bodies thought experiment within a scientific controversy and relied on this thought experiment — along with several other resources — *as a basis for discussion* of a question that must have sounded "meaningless" to peripatetic scholars, say, "can bodies (of different weights) move in a void (with equal speeds)?"

Now, there is another important aspect of Galileo's approach to persuasion which I wish to draw attention to in connection with the legend of the Leaning Tower of Pisa.

Galileo's culture of experiment and its role in his research program

Galileo's culture of experiment included initial experiments, thought experiments, and public demonstrations. He is known to have demonstrated his findings and held public exhibits such as the one at the top of the Tower of San Marco in Venice as we have seen. In fact, Paul Feyerabend tried to show in (2010 [1970]) that Galileo's research program had an important propaganda component. I shall not comment Galileo's culture of experiment as a topic in general but focus on its central connection with the heuristic function of his falling-bodies thought experiment.

I pose the following questions: how could Galileo confirm his hypothesis on the existence of a vacuum and show its key property on motion? More specifically, how could Galileo *remove the medium* to actually be able to reproduce his thought experiment in conditions similar to those of running it from the top of the Leaning Tower of Pisa? It turns out that this would require significant developments in vacuum science and engineering. Now, what is the relation between the thought experiment itself and the developments which could one day actually enable running it as an actual experiment? Note that this a sensible question in view of the Turing test dilemma.

As known, Galileo's falling-bodies thought experiment and related initiatives of his formed a tradition of making science demonstrations to authorities and the general public. Also commenting on Galileo's legacy relative to science propaganda, Robert Crease made this most interesting comment which I quote in connection with my point:

> Yet another fascinating side to Galileo's experiments is the way that they slowly transformed from genuine scientific inquiries into public displays. After Galileo's death, scientists including Robert Boyle and Willem's Gravesande built air pumps and special chambers to explore vertical fall in evacuated environments. King George III, for instance, once witnessed a demonstration involving a feather and a one-guinea coin falling together inside an evacuated tube. The popularity of such demonstrations continues to this day, featuring in many hands-on science exhibits. (CREASE, 2003)

Crease's side note hit right on target the line of thought I wish to suggest. It turns out that the existence of a vacuum and the confirmation of its properties requires the development of vacuum science and engineering. In particular, the *removal of the medium* as idealized in Galileo's leaning-tower thought experiment requires the science and technology of large-scale vacuum chambers. In sum, there is a research and development path all the way from the "air pumps and special chambers" built by figures such as Robert Boyle (1627-1692) to the space program that would eventually enable us to land on the Moon. And Galileo's falling-bodies thought experiment may well be acknowledged as one of its starting points. In fact, Boyle reported in his autobiographical notes that he was once in his teens visiting Florence, "reading widely in the sciences, particularly the new celestial system of Copernicus and his followers" just in early January 1642 when Galileo died as an exiled prisoner of the Church (FULTON, 1960, p. 119). The event is said to have "profoundly stirred the impressionable young Englishman" who later wrote that at Florence he had encountered "the new paradoxes of the great star-gazer Galileo." Boyle also commented on Galilleo's affair with the Church. He thought very highly of Galileo (*Ibid.*). What exact influence did Galileo have on Boyle is beyond my point here. In any case, the small-scale experiments of Boyle were able to demonstrate the existence of motion in a void. He achieved that by resorting to the culture of science demonstrations as exemplified by Galileo, say, in his showcases of sky-gazing through his telescopes.

And yet the Leaning Tower of Pisa in particular is 57-meters high, which is the order of magnitude required for the study of free-falling bodies in the context of Galileo's dispute with the peripatetics. The idealized conditions of Galileo's falling-bodies thought experiment as it was could only be reproduced after the advent of large-scale vacuum chambers. To my knowledge the first such chamber to have been built was NASA so-called Space Power Facility, with 37-meters height, in 1969 in Ohio. Suffice to say that related events can be seen as having confirmed that, in a vacuum, Galileo's thought experiment would render the acclaimed outcome. Bodies of different weights fall alike, indeed. I refer to these events later (§3.10).

Now, let us recall Gandy's anecdote about the intended function(s) of the Turing test as quoted at the beginning of this chapter. He related that Turing's 1950 paper "was intended not so much as a penetrating contribution to philosophy but as propaganda" (1996, p. 125). Also, Marvin Minsky interpreted that the imitation game was sort of a joke. But if so, then, as we have seen, it was a joke packed with a scientific hypothesis inside. So, in a nutshell, what can we take from all this to inform our view of Turing's imitation game?

Turing's thought experiment in light of Galileo's

Both Turing and Galileo's thought experiments were imagined and presented within a scientific controversy. Both thought experiments performed critical and heuristic functions. They were a basis for criticizing key assumptions of their intellectual opponents, but they did not refute those assumptions. They invited us to imagine a previously unthinkable situation in which key

idealizations were present. Galileo idealized motion in a vacuum. Turing idealized a digital computer whose storage capacity was orders of magnitude larger than that of the machines of his day, and whose program would enable them to learn from experience in general like a child brain. The heuristic function of Galileo's thought experiment was to represent his existential hypothesis on motion in a vacuum. The heuristic function of Turing's was to represent his existential hypothesis on thinking machines. The feasibility condition for Galileo's thought experiment to be run as an actual experiment to decide for his hypothesis was the development of the science and engineering of a large-scale vacuum chamber. It came by in the last century. A decision about the outcome of Galileo's thought experiment is whether or not bodies of different weights fall alike. Now, if there still was any reasonable doubt about this result, since the experiment became feasible there should be no more. In a vacuum, they do. The feasibility condition for Turing's to work the same way is the development of the science and engineering of machine learning. This enterprise started in the early 1950's and had some new developments most recently. Will it lead to make the idealized condition of a child-brain program feasible? Perhaps, if Turing's more specific vision can recruit. The next decades or century shall inform. A decision about the outcome of Turing's thought experiment is whether or not a mature child-brain program which learned from general experience deceives an average interrogator and passes the test.

It is interesting to note at this point is that there has been much debate about whether or not Turing and Galileo's thought experiments could be taken as decisive. Nonetheless, with the Turing test dilemma in mind, let us consider the two stories that have grown around Galileo's famous thought experiment as we have seen, the philosophy of science story and the history of science story. Now, if there is a lesson to be learned from them, it may be the following:

The elusive decision problem of Galileo's falling-bodies thought experiment. A decision about the empirical question at issue in Galileo's thought experiment could only come about when the very science it was meant to vindicate was so much developed that its original contention had actually become trivial. By that time, there was hardly any point anymore in conducting it other than paying honors.

I think that the reason for the elusiveness of this problem is the presence of an existential hypothesis, which renders, in case the thought experiment is seen as an actual experiment, a *petitio principii*. As we have seen, Norton drew attention to this problem. He remarked: "Salviati is in no position to assume away the medium." And then completed: "[t]o assume the possibility of a vacuum at this point would be to beg the main question under discussion" (1996, p. 344). In the words of Hayes and Ford: "The [imitation] tests are circular: they define the qualities they are claiming to be evidence for" (1995, p. 974).

In light of this, we may now revisit Turing's outline of his beliefs:

> The original question, 'Can machines think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one

> will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs. The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any unproved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures, no harm can result. Conjectures are of great importance since they suggest useful lines of research. (TURING, 1950, p. 442)

It is as if Turing understood that he was discussing an existential hypothesis, and that the problem had an important social and cultural, and even a propaganda component. Paul Feyerabend (2010 [1970]) made in my view a sensible related remark (cf. p. 72, note 22) in his reply to a 1986 criticism he received by Alan Chalmers. In fact, we may now remind of Gandy's anecdote about the intended function(s) of the Turing test as quoted at the beginning of this chapter. Given all that we have seen so far, his anecdote seems right on target as an explanation of Turing's motive to come up with the imitation game. I shall now return to Crease to conclude:

> Galileo played a seminal role in transforming that framework [of everyday experience], in developing the abstract thinking involved in the new one, and in illustrating its importance. So what if there was no original [leaning tower] experiment? Galileo inspired an entire genre of experiments and demonstrations that allow us to change how we think and see. We might as well refer to these as the offspring of Galileo's experiment at the Leaning Tower of Pisa. (CREASE, 2003)

I hope to have succeeded in showing that Turing's imitation game is a legitimate offspring of Galileo's leaning tower thought experiment. Unlike the latter, the empirical status of the former is yet in the open. I believe, however, that the analogy is strong enough to grant the imitation game to be classed as a thought experiment in science.

3.10 Epilogue

Galileo's falling-bodies thought experiment has caught the imagination of many over the centuries. In this connection, besides his sensible reference to Boyle's air pumps, Robert Crease (2003) pointed to an event that took place on the lunar surface. Apollo 15 is reported to have been the fourth crewed landing on the Moon, on 26 July 1971. A few days later on 2 August 1971 commander David R. Scott made a symbolic demonstration of Galileo's thought experiment from the lunar surface. He used a feather and a hammer, and the crew videotaped the demonstration.[5] Crease thus transcribed Scott's narrative during the demonstration from the lunar-surface:

> SCOTT: Well, in my left hand I have a feather; in my right hand, a hammer. And I guess one of the reasons we got here today was because of a gentleman named Galileo, a long time ago, who made a rather significant discovery about falling objects in gravity fields. And we thought: 'Where would be a better

[5] The Apollo 15 video is available at <http://nssdc.gsfc.nasa.gov/planetary/lunar/apollo_15_feather_drop.html>. Access on 20 Nov. 2020.

> place to confirm his findings than on the Moon?'.
>
> [Camera zooms in on Scott's hands. One is holding a feather, the other a hammer. The camera pulls back to show the Falcon – the Apollo 15 landing craft – and the lunar horizon.]
>
> SCOTT: And so we thought we'd try it here for you. The feather happens to be, appropriately, a falcon feather for our Falcon. And I'll drop the two of them here and, hopefully, they'll hit the ground at the same time.
>
> [Scott releases hammer and feather. They hit the ground at about the same time.]
>
> SCOTT: How about that! Mr Galileo was correct in his findings.
> (CREASE, 2003)

Without being subject to the atmosphere of the Earth, the bodies fell alike. Recently, a variant of Galileo's thought experiment has been reproduced in realistic conditions at NASA Space Power Facility in Cleveland,[6] Ohio, by initiative of the the BBC. (By chance the same institution that recorded three oral communications of Turing.) A feather and a bowling ball were dropped from nearly the top of the 37-meters high vacuum chamber. In a first trial, the chamber was naturally filled with air. The feather fell mush slower than the ball. In a second trial, the air was removed from the chamber. They fell alike, and in fact hit the ground exactly at the same time.[7]

Naomi Araki is a fellow researcher of Turing's question (2018). She noted that science-fiction writer Arthur Clarke addressed it in his famous 1968 novel *2001: a space odyssey*:

> The sixth member of the crew cared for none of these things, for it was not human. It was the highly advanced HAL 9000 computer, the brain and nervous system of the ship. [...] Whether HAL [Heuristically programmed ALgorithmic computer] could actually think was a question which had been settled by the British mathematician Alan Turing back in the 1940s. Turing had pointed out that, if one could carry out a prolonged conversation with a machine — whether by typewriter or microphone was immaterial — without being able to distinguish between its replies and those that a man might give, then the machine was thinking, by any sensible definition of the word. HAL could pass the Turing test with ease. (CLARKE, 1968, p. 97)

We see that in fiction Turing's thought experiment also meets with Galileo's. It remains to be seen whether this encounter will come to reality.

3.11 Chapter acknowledgements

I am very grateful to Prof. Pío Garcia to have pointed out to me in June 2018, in the occasion of the XI Encuentro de Filosofía e Historia de la Ciencia del Cono Sur in Buenos Aires, Argentina, the importance of Jefferson's 1949 paper to understand Turing's 1950 test. I would also like to thank Prof. Garcia to have in December 2019 suggested me to do a thorough review of the

[6] See <http://www.nasa.gov/multimedia/imagegallery/image_feature_648.html>. Access on 20 Nov. 2020.
[7] Available at: <http://www.youtube.com/watch?v=E43-CfukEgs>. Access on 20 Nov. 2020.

literature on thought experiments in science including Ernst Mach's analysis and the Norton-Brown controversy. It should be obvious that the importance of Prof. Garcia's contributions to this dissertation are invaluable. Possible mistakes in my elaborations out of these materials are of course my sole responsibility.

I would like to thank Plínio Smith for his teachings about philosophical skepticism, about the Pyrronist notion and use of dialectic argument in the context of ancient philosophy. I also thank Lucas Petroni for the suggestion of looking at how Bertrand Russell introduced the notion of dialectics in his *History*.

Conclusion

In this dissertation I have presented a study of Alan Turing's imitation game or test and its central question, whether machines can think. At the end of each chapter I have consolidated the key points developed in connection with the specific problem of each chapter. I shall now briefly recapitulate the key points developed in general, essentially revisiting the Introduction.

I have examined the historical and epistemological roots of Turing's various versions of imitation game or test and argued that they came out from within a dialogue. More specifically, it was born out of a scientific controversy (1949-1952), most notably with physicist and computer pioneer Douglas Hartree, chemist and philosopher Michael Polanyi, and neurosurgeon Geoffrey Jefferson. I hope to have succeeded in showing that the study of this controversy is crucial to understand Turing's 1950 proposal and imitation game in depth. And yet it has barely been noticed in connection with the famous Turing test in the analytic philosophy literature. This can be verified by looking at, say, *Stanford Encyclopedia of Philosophy*'s dedicated entry, or the most recent of various dedicated special issues on the *Minds and Machines* journal which is about to appear. Besides the irregular availability of sources over time, this neglect of the origins of Turing's imitation tests (1948-1952) even when the sources are available may help explain the sheer heterogeneity that we find still today in the secondary literature about the Turing test. Overall, I have strived to place Turing's views in their historical, social and cultural context, which can shed new light at several aspects of the so-called Turing test, including what I have called the Turing test dilemma. I have collected evidence to reclaim the scientific and philosophical value of Turing's 1950 proposal for the sake of the discussion in the years to come.

I have formulated and addressed three main philosophical problems. In connection with my study of Turing's 1950 proposal (§3), I have examined Turing's profile (§1) and deeper views (§2). Let us now collect the key points developed in each of theses efforts.

At this point Turing has been widely identified as a war hero, as a brilliant mathematician and scientist, but not yet as clearly as a philosopher. He challenged the conventional wisdom of what machines really were or could be, and suggested releasing them from their duties as slaves. For that he received antagonism also at a personal level, even if but subtly, for the non-conformist stripe of his views in Cold War Britain. He has been accused of Promethean irreverence and been associated, among other things, with the image of Dr. Frankenstein. The first problem I have addressed was the identification of Turing's specific (Promethean) ambition which led him to announce that machines will think. Turing thought most clearly, I have interpreted, that digital computers were scientific instruments, say, like Galileo's telescopes. They would do great service for the study of the human mindbrain. Also, a reader of Victorian novelist Samuel Butler, Turing prophesized a future world pervaded by intelligent machines which may be seen

as a dystopia just as much as a utopia. Turing drew attention to the advent of superintelligent machines, or the possibility of machines to outstrip our intellectual powers and take control. For him, although it was not a certain event, its possibility was true. And this poses the question: if he thought so and still wanted to build intelligent machines, does that mean that Turing collaborated towards a dystopian future? I answered no, and argued that he was rather raising a precautionary voice. In the presence of the crass chauvinism of some of his contenders, however, he could not but establish epistemological relations between the possibility of *intelligent* machines and the refutation of what he saw as species and gender biases of his contenders. For this reason, I think, Turing's prophesized future pervaded by intelligent machines may be seen as a dystopia just as much as a utopia. It depends on the social and cultural background of who is seeing. *Superintelligent* machines were for him a true but distant possibility that pushed his argument to the limit from both epistemological and social stances. Turing also expressed a dislike towards the dismissal of the ontological distinction between the natural and the artificial. The evidence I have collected is unfavorable to associating Turing with, say, modern transhumanist movements.

In the second problem, over and above the mere proposal of a test for machine intelligence based on verbal behavior, I have studied what views of thinking, intelligence, the mind and the brain Turing actually posed. In this connection I have singled out and discussed key passages from Turing sources all the way from the outset of his concept of machine intelligence to its endgame (1936-1952). I have studied Turing's proposition "machines can think" and its implied existential hypothesis — "there exists (will exist) a thinking machine" — from a point of view of the history of the philosophy of science. Contrary to common readings of Turing, I found that Turing held a non-obvious realist (and physicalist) attitude towards the existence of a mechanical mindbrain which he conjectured to frame the human and whose digital replica he intended to build in the machine. Also, I have examined Turing's views through the lens of a clear-cut distinction between Turing's epistemology and his ontology of thinking or intelligence. I have found that the former is based on causing wonder on an external observer. It is subjectivist and emotional, yet observable. The latter is based on learning like a brain from experience. It is objectivist and mathematical, yet only indirectly observable. Contrary to common readings of Turing, and specially those that construe him as a behaviorist, I found that Turing held a realist (and physicalist) attitude towards the existence of a mechanical mindbrain which he conjectured to frame the human and whose digital replica he intended to build in the machine.

Turing's 1950 paper has been acknowledged as a complex and multi-layered text. We have seen that the interpretation of Turing's 1950 proposal is still remarkably controversial. Two opposing — and in fact incommensurable — views exist in the secondary literature concerning the question on whether or not Turing proposed his imitation test as an actual experiment to decide for machine intelligence. And yet, even among those that saw in it the proposal of an actual — definite and practical — scientific experiment, further heterogeneity arises, for there are supporters *and* critics of the significance of Turing's test as such. Most supporters in this class either dismissed or shrank the element of gender imitation in the test. Most critics in

the same class contended that Turing proposed an operational definition of intelligence and a form of behaviorism that is reductive of the human mind. Others yet have seen in Turing's proposal not an actual experiment at all but either just "a joke" or at most a historical manifesto for artificial intelligence with no scientific content inside. Some interpreters in this latter class acknowledged gender imitation as an important element in the structure of the test but considered that it rather testifies against its seriousness. I have referred to this philosophical quagmire as the Turing test dilemma and have addressed it as my third problem. I have found that a rigorous exegesis of Turing's 1950 text, together with knowledge of a series of related historical events — crucially, the 1949 Manchester seminars (§A.4.2) — and the identification of Turing's proposition "machines can think" as implying an existential hypothesis, altogether, is suggestive that Turing cannot not have proposed the imitation game as something other than a thought experiment, no matter he was conscious about that or not. For an example, the controversial gender question, I have found, plays a key function in the thought experiment which is to encode in wit a rebuttal to a serious argument posed by Jefferson in his 1949 Lister Oration about sex hormones and the physiology of behavior. Thought experiments in science may be feasible to be run and this seems to be the case of Turing's, and yet that does not mean that running the imitation game is a sensible scientific project. I have drawn on Ernst Mach's analysis of the nature of thought experiments in science, and also on the history of the legend of the Leaning Tower of Pisa. Galileo's thought experiment, I have held, informs the current situation concerning Turing's.

So, we have seen overall a wealth of historical, social and cultural elements of that controversy and how these may have influenced Turing's moves. But we can also observe that Turing established a longer-term dialogue within the history of philosophy. He suggested the existence of a mechanism which could be encoded in digital computers and, however different than human thinking in nature, would be indistinguishable from it when it comes to speech. Turing thus committed, as known and as we have seen, to a much earlier proposal of René Descartes's which implied unrestricted conversation as a proxy for thinking. Descartes presumed that a true talking machine was not practically possible to exist, while for Turing this was a matter of allowing for some tractable advances in the science and technology of modern computing to take place. For both, the actual existence of a thinking machine was rather an empirical claim. We have seen that Turing met precisely the terms required by Descartes, and actually improved on it from the point of view of an experiment design. And he scientifically challenged the cartesian explanation of reason as the unique belonging and citadel of a rational soul. By questioning any form of chauvinism, be it of species or gender whatsoever, Turing saw intelligence instead as essentially a result of learning like a brain from experience. If granted enough memory and a suitable program, Turing posited, machines would be able to learn for themselves from experience in general, with no need to be turned off and reset before they could react to a newly posed challenge. He said (§2.5) that this "leaves open the question as to whether we will or will not eventually succeed in finding such a programme" (p. 484). "I, personally," he completed, "am inclined to believe that such a programme will be found." Whether Turing's hypothesis will

be confirmed or not may well be, in effect, one of the key scientific questions of this century.

Bibliography

AAMOTH, D. Interview with Eugene Goostman, the fake kid who passed the Turing test. *Time*, n. 9 jun., 2014. Available at: <http://time.com/2847900/eugene-goostman-turing-test/>. Access on 20 Nov. 2020. Cited in page 119.

ABRAMSON, D. Turing's responses to two objections. *Minds and Machines*, v. 18, n. 2, p. 147–67, 2008. doi: 10.1007/s11023-008-9094-6. Cited in page 64.

ABRAMSON, D. Descartes' influence on Turing. *Studies in History and Philosophy of Science: Part A*, v. 42, n. 4, p. 544–51, 2011. Doi:10.1016/j.shpsa.2011.09.004. Cited 7 times in pages 64, 91, 98, 103, 123, 149, and 154.

ADLER, C. G.; COULTER, B. L. Galileo and the Tower of Pisa experiment. *American Journal of Physics*, v. 46, n. 3, p. 199–201, 1978. doi: 10.1119/1.11165. Cited 2 times in pages 186 and 187.

AGAR, J. *Turing and the universal machine: the making of the modern computer*. London: Icon Books, 2001. Cited in page 16.

ANDERSON, D. Max Newman: topologist, codebreaker, and pioneer of computing. *IEEE Annals of the History of Computing*, v. 29, n. 3, p. 76–81, 2007. doi: 10.1109/MAHC.2007.42. Cited in page 243.

ANDERSON, D. Max Newman: forgotten man of early British computing. *Communications of the ACM*, v. 56, n. 5, p. 29–31, 2013. doi: 10.1145/2447976.2447986. Cited 2 times in pages 221 and 241.

ARAKI, N. Alan turing's question. *Bulletin of Hiroshima Institute of Technology*, v. 52, p. 33–42, 2018. Cited in page 192.

ARTHUR, R. On thought experiments as a priori science. *International Studies in the Philosophy of Science*, v. 13, n. 3, p. 215–29, 1999. doi: 10.1080/02698599908573622. Cited in page 181.

BABBAGE, C. *The Ninth Bridgewater Treatise: a fragment*. Cambridge: Cambridge University Press, 2009 [1837]. (Cambridge Library Collection: Religion). Cited in page 51.

BBC. Government rejects pardon request for Alan Turing. *BBC News*, 8 Mar. 2012. Available at: <http://www.bbc.com/news/technology-16919012>. Access on 20 July 2020. Cited in page 52.

BBC. Royal pardon for codebreaker Alan Turing. *BBC News*, 24 Dec. 2013. Available at: <http://www.bbc.com/news/technology-25495315>. Access on 20 July 2020. Cited in page 52.

BEDINI, S. A. The role of automata in the history of technology. *Technology and Culture*, v. 5, n. 1, p. 24–42, 1964. doi: 10.2307/3101120. Cited in page 42.

BLOCK, N. Psychologism and behaviorism. *The Philosophical Review*, XC, n. 1, p. 5–43, 1981. Doi: 10.2307/2184371. Cited in page 122.

BLUM, P. R. Michael Polanyi: can the mind be represented by a machine? documents of the discussion in 1949. *Polanyiana*, v. 19, n. 1-2, p. 35–60, 2010. Available at: <http://www.polanyi.bme.hu/folyoirat/2010-01/2010-1-2-03-Blum.pdf>. Access on 3 Jul. 2020. Cited 5 times in pages 96, 146, 252, 273, and 274.

BMJ. Mind, machine, and man. *British Medical Journal*, v. 1, n. 4616, p. 1129–1130, 1949. Cited 5 times in pages 29, 40, 47, 248, and 250.

BOGUE, J. Y. *Christmas greetings letter from J. Yule Bogue sent to Warren McCulloch in* c. *December 1949*. 1949. Found and transcribed by Jonathan Swinton. Original is kept in the MIT American Philosophical Society Warren McCulloch archive. Facsimile available at: <http://www.manturing.net/manufacturing-blog/2019/6/3/manchester-minds-and-mit-ones>. Access on 10 April 2020. Cited 3 times in pages 148, 255, and 256.

BOWDEN, B. V. (Ed.). *Faster than Thought: a symposium on digital computing machines*. London: Pitman, 1953. Facsimile available at: <http://archive.org/details/fasterthanthough00bvbo>. Cited in page 274.

BROWN, J. R. Thought experiments since the scientific revolution. *International Studies in the Philosophy of Science*, v. 1, n. 1, p. 1–15, 1986. doi: 10.1080/02698598608573279. Cited in page 181.

BROWN, J. R. *The laboratory of the mind: thought experiments in the natural sciences*. London: Routledge, 1991. (Philosophical Issues in Science). Cited 7 times in pages 112, 113, 152, 163, 164, 181, and 187.

BUTLER, S. *Erewhon or over the range*. London: Trubner & Co., 1872. Cited 2 times in pages 38 and 39.

CARNAP, R. *Der logische Aufbau der Welt*. Third. Berlin: Felix Meiner Verlag, 1966 [1928]. Cited in page 84.

CHAPLIN, C. *My Autobiography*. Pocket books. New York: Simon & Schuster, 1966 [1964]. Cited in page 48.

CHOMSKY, N. Language and nature. *Mind*, v. 104, n. 413, p. 1–61, 1995. doi: 10.1093/mind/104.413.1. Cited 2 times in pages 116 and 117.

CHRISTENSEN, R. S. et al. Laboratory test of the Galilean universality of the free fall experiment. *Physics Education*, v. 49, n. 2, p. 201–10, 2014. doi: 10.1088/0031-9120/49/2/201. Cited in page 187.

CHURCH, A. Review: A. M. Turing, On computable numbers, with an application to the Entscheidungsproblem. *Journal of Symbolic Logic*, v. 2, n. 1, p. 42–3, 1937. Cited 2 times in pages 35 and 217.

CHURCHILL, W. *The Sinews of Peace ('Iron Curtain Speech')*. 1946. Wiston Churchill Speeches. International Churchill Society. Available at: <http://winstonchurchill.org/resources/speeches/1946-1963-elder-statesman/the-sinews-of-peace/>. Access on 20 July 2020. Cited in page 48.

CLARKE, A. C. *2001: a space odyssey*. New York: Signet, 1968. Cited in page 192.

COOPER, L. *Aristotle, Galileo, and the Tower of Pisa*. New York: Ithaca, 1935. Cited in page 185.

COPELAND, B. J. *Artificial intelligence: a philosophical introduction*. New Jersey: Wiley-Blackwell, 1993. Cited in page 159.

COPELAND, B. J. A lecture and two radio broadcasts on machine intelligence by Alan Turing. In: FURUKAWA, K.; MICHIE, D.; MUGGLETON, S. (Ed.). *Machine Intelligence 15*. Oxford: Oxford University Press, 1999. Cited 4 times in pages 14, 258, 260, and 263.

COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004. Cited 24 times in pages 38, 72, 97, 214, 219, 220, 224, 225, 226, 230, 231, 236, 237, 238, 241, 242, 243, 244, 256, 258, 260, 263, 267, and 272.

COPELAND, B. J. Colossus and the rise of the modern computer. In: COPELAND, B. J. (Ed.). *Colossus: the secrets of Bletchley Park's codebreaking computers*. Oxford: Oxford University Press, 2006. cap. 9, p. 101–15. Cited 3 times in pages 241, 242, and 243.

COPELAND, B. J. The Manchester computer: A revised history. *IEEE Annals of the History of Computing*, v. 33, n. 1, p. 4–37, 2011. Cited 3 times in pages 150, 243, and 247.

COPELAND, B. J. *Turing: pioneer of the Information Age*. Oxford: Oxford University Press, 2012. Cited 7 times in pages 16, 23, 29, 48, 52, 229, and 231.

COPELAND, B. J. Bombes. In: COPELAND, B. J. et al. (Ed.). *The Turing Guide*. Oxford: Oxford University Press, 2017. cap. 12, p. 109–28. Cited in page 230.

COPELAND, B. J. Intelligent machinery. In: COPELAND, B. J. et al. (Ed.). *The Turing Guide*. Oxford: Oxford University Press, 2017. cap. 25, p. 265–75. Cited 2 times in pages 229 and 230.

COPELAND, B. J. et al. *Alan Turing's Automatic Computing Engine: the master codebreaker's struggle to build the modern computer*. Oxford: Oxford University Press, 2005. Cited in page 237.

COPELAND, B. J.; PRINZ, D. Computer chess–the first moments. In: COPELAND, B. J. et al. (Ed.). *The Turing Guide*. Oxford: Oxford University Press, 2017. cap. 31, p. 327–46. Cited 3 times in pages 230, 244, and 245.

COPELAND, B. J.; PROUDFOOT, D. *On Alan Turing's anticipation of connectionism*. [S.l.: s.n.], 1996. v. 108. 361-77 p. doi: 10.1007/BF00413694. Cited 2 times in pages 94 and 244.

COPELAND, B. J.; PROUDFOOT, D. Alan Turing: father of the modern computer. *The Rutherford Journal*, v. 4, 2012. Web-book Special Issue for the 2012 Alan Turing Centenary Year. Available at: <http://www.rutherfordjournal.org/article040101.html>. Access on 20 July 2020. Cited in page 241.

COPELAND, J. Narrow versus wide mechanism: including a re-examination of Turing's views on the mind-machine issue. *The Journal of Philosophy*, v. 97, n. 1, p. 5–32, 2000. doi: 10.2307/2678472. Cited 2 times in pages 64 and 91.

COPELAND, J. The Turing test. *Minds and Machines*, v. 10, n. 4, p. 519–39, 2000. doi: 10.1023/A:1011285919106. Cited 6 times in pages 63, 91, 98, 103, 123, and 124.

COPELAND, J. Turing and the physics of the mind. In: COOPER, S. B.; van Leeuwen, J. (Ed.). *Alan Turing: his work and impact*. Amsterdam: Elsevier Science, 2013. Cited in page 64.

COPELAND, J. Crime and punishment. In: COPELAND, B. J. et al. (Ed.). *The Turing Guide*. Oxford: Oxford University Press, 2017. cap. 4, p. 35–40. Cited in page 49.

COTTINGHAM, J. Cartesian dualism: theology, metaphysics and science. In: COTTINGHAM, J. (Ed.). *The Cambridge companion to Descartes*. Cambridge: Cambridge University Press, 1992, (Cambridge companions to philosophy). cap. 8, p. 236–57. Cited 2 times in pages 12 and 169.

CREASE, R. P. The legend of the leaning tower. *Physics World*, n. 4 Feb., 2003. Available at: <http://physicsworld.com/a/the-legend-of-the-leaning-tower/>. Access on 6 Oct. 2020. Cited 3 times in pages 188, 191, and 192.

DARWIN, C. *Letter from Darwin to Appleton*. Teddington, 1947. Facsimile available at: <http://www.alanturing.net/darwin_appleton_23jul47/>. Access on 20 July 2020. Cited 2 times in pages 73 and 241.

DARWIN, C. *NPL executive committee minutes 28 Sept. 1948*. Teddington, 1948. Facsimile available at <http://www.alanturing.net/turing_archive/archive/l/l100/l100.php>. Access on 20 Jul. 2020. Cited in page 244.

DARWIN, C. G. Douglas Rayner Hartree, 1897-1958. *Biographical memoirs of fellows of the Royal Society*, v. 4, n. Nov., p. 102–16, 1958. doi: 10.1098/rsbm.1958.0010. Cited in page 143.

DAVIES, C. PM's apology to codebreaker Alan Turing: we were inhumane. *The Guardian*, 11 Sep. 2009. Available at: <http://www.theguardian.com/world/2009/sep/11/pm-apology-to-alan-turing>. Access on 20 July 2020. Cited in page 52.

DAVIS, M. (Ed.). *The undecidable: basic papers on undecidable propositions, unsolvable problems and computable functions*. New York: Raven, 1965. Cited in page 222.

DAVIS, M. How subtle is Gödel's theorem? more on Roger Penrose. *Behavioral and Brain Sciences*, v. 16, n. 3, p. 611–2, 1993. doi: 10.1017/S0140525X00031915. Cited in page 226.

DAVIS, M. *The universal computer: the road from Leibniz to Turing*. New York: W. W. Norton, 2000. Cited in page 221.

DENNETT, D. Can machines think? In: TEUSCHER, C. (Ed.). *Alan Turing: Life and Legacy of a Great Thinker*. Berlin: Springer, 2006 [1984]. p. 295–316. Reprinted from (Ed.) M. G. Shafto, *How we know*, (San Francisco: Harper & Row) 121-145, 1984, plus postscripts "Eyes, ears, hands and history" (1985) and (1997, no title). Cited 6 times in pages 16, 98, 110, 111, 123, and 125.

DENNETT, D. Postscript (1997, no title) to "Can machines think?" (1984). In: TEUSCHER, C. (Ed.). *Alan Turing: Life and Legacy of a Great Thinker*. Berlin: Springer, 2006 [1997]. p. 314–6. Cited 3 times in pages 112, 120, and 125.

DENNETT, D. *Intuition pumps and other tools for thinking*. New York: W. W. Norton & Company, 2013. Cited 5 times in pages 50, 112, 113, 114, and 180.

DESCARTES, R. *The philosophical writings of Descartes*. Cambridge: Cambridge University Press, 1985. I. Trans. and ed. by John Cottingham, Robert Stoothoff and Dugald Murdoch. Cited 2 times in pages 160 and 169.

DESCARTES, R. Discourse and essays. In: COTTINGHAM, J.; STOOTHOFF, R.; MURDOCH, D. (Ed.). *The philosophical writings of Descartes*. Cambridge: Cambridge University Press, 1985 [1637]. Vol. I, p. 111–51. Cited 2 times in pages 12 and 154.

DESCARTES, R. Letter to Mersenne, end of November 1633. In: COTTINGHAM, J. et al. (Ed.). *The philosophical writings of Descartes*. Cambridge: Cambridge University Press, 1991 [1633]. Vol. III: the correspondence, p. 40–1. Trans. and ed. by John Cottingham, Robert Stoothoff, Dugald Murdoch and Anthony Kenny. Cited in page 13.

DESCARTES, R. To More, 5 february 1649. In: COTTINGHAM, J. et al. (Ed.). *The philosophical writings of Descartes*. Cambridge: Cambridge University Press, 1991 [1649]. III: the correspondence, p. 360–7. Cited in page 160.

DIJKSTRA, E. *The threats to computing science*. 1984. Talk delivered at the ACM 1984 South Central Regional Conference, November 16–18, Austin, Texas. Available at: http://www.cs.utexas.edu/users/EWD/transcriptions/EWD08xx/EWD898.html. Access on: 20 Jun 2020. Cited in page 116.

DRAKE, S. *Galileo at work: his scientific biography*. Chicago: The University of Chicago Press, 1978. Cited 2 times in pages 184 and 185.

EPSTEIN, R. The quest for the thinking computer. *AI Magazine*, v. 13, n. 2, p. 81–95, 1992. doi: 10.1609/aimag.v13i2.993. Cited in page 111.

ERION, G. J. The Cartesian test for automatism. *Minds and Machines*, v. 11, n. 1, p. 29–39, 2001. Cited in page 160.

EVANS, C. R.; ROBERTSON, A. D. J. (Ed.). *Key papers: Cybernetics*. London: Butterworths, 1968. Cited in page 14.

FAHIE, J. *Memorials of Galileo (1564–1642)*. London: The Courier Press, 1929. Cited in page 55.

FEFERMAN, S. Ordinal logics. *Routledge Encyclopedia of Philosophy*, 1998. doi: 10.4324/9780415249126-Y012-1. Cited in page 227.

FEIGL, H. Existential hypotheses: realistic versus phenomenalistic interpretations. *Philosophy of Science*, v. 17, n. 1, p. 35–62, 1950. Cited 7 times in pages 20, 83, 84, 86, 89, 95, and 138.

FEIGL, H. *The "mental" and the "physical": the essay and a postscript*. Minneapolis: University of Minnesota Press, 1967 [1958]. Cited in page 95.

FEINBERG, G. Fall of bodies near the earth. *American Journal of Physics*, v. 33, n. 6, p. 501–3, 1965. doi: 10.1119/1.1971740. Cited in page 186.

FEYERABEND, P. *Against method*. Fourth. London: Verso, 2010 [1970]. Cited 2 times in pages 188 and 191.

FINOCCHIARO, M. A. *The Galileo affair: a documentary history*. California: University of California Press, 1989. (California Studies in the History of Science). Cited in page 50.

FINOCCHIARO, M. A. *Retrying Galileo, 1633-1992*. Berkeley: University of California Press, 2005. Cited in page 55.

FRENCH, R. Subcognition and the limits of the Turing test. *Mind*, XCIX, n. 393, p. 53–65, 1990. doi: 10.1093/mind/XCIX.393.53. Cited in page 120.

FROMM, E. Prophets and priests. In: *On disobedience: why freedom means saying "no" to power*. Harper perennial modern thought. New York: Harper & Row, 2010 [1967]. cap. II, p. 13–40. Reprinted from (Ed.) Ralph Schoenman, *Bertrand Russell, philosopher of the century: essays in his honour*, (London: Allen & Unwin, 1967). Cited in page 52.

FRYER, D. M.; MARSHALL, J. C. The motives of Jacques de Vaucanson. *Technology and Culture*, v. 20, n. 2, p. 257–69, 1979. Doi: 10.2307/3103866. Cited 2 times in pages 42 and 119.

FULTON, J. F. The honourable Robert Boyle, F. R. S. (1627-1692). *Notes and Records of the Royal Society*, v. 15, n. 1, 1960. doi: 10.1098/rsnr.1960.0012. Cited in page 189.

GALILEI, G. Excerpts from *The assayer*. In: DRAKE, S. (Ed.). *Discoveries and Opinions of Galileo*. 24th edition. ed. Norwell, MA: Anchor, 1957 [1623]. p. 229–80. Cited in page 51.

GALILEI, G. De motu (c. 1590) (translated with introduction and notes by. In: DRABKIN, I. E.; DRAKE, S. (Ed.). *On Motion and On Mechanics*. Madison: University of Wisconsin Press, 1960 [1590]. Cited in page 185.

GALILEI, G. *Two new sciences*. Madison: University of Wisconsin Press, 1974 [1638]. Translated by Stillman Drake. Cited 6 times in pages 181, 183, 185, 186, 187, and 188.

GANDY, R. Human versus mechanical intelligence. In: MILLICAN, P.; CLARK, A. (Ed.). *Machines and Thought: The Legacy of Alan Turing*. Oxford: Oxford University Press, 1996. v. 1. Cited 4 times in pages 106, 148, 189, and 256.

GENDLER, T. S. Galileo and the indispensability of scientific thought experiment. *British Journal for the Philosophy of Science*, v. 49, n. 3, p. 397–424, 1998. doi: 10.1093/bjps/49.3.397. Cited in page 181.

GENOVA, J. Turing's sexual guessing game. *Social Epistemology*, v. 8, n. 4, p. 13–26, 1994. doi: 10.1080/02691729408578758. Cited 7 times in pages 62, 63, 90, 98, 103, 133, and 134.

GÖDEL, K. On formally undecidable propositions of Principia Mathematica and related Systems I. In: DAVIS, M. (Ed.). *The undecidable: basic papers on undecidable propositions, unsolvable problems and computable functions*. New York: Raven, 1965 [1931]. p. 5–38. English translation of K. Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatshefte für Mathematik und Physik* 38: 173-98, 1931. Cited 2 times in pages 222 and 226.

GOODING, D. C. Thought experiments. In: CRAIG, E. (Ed.). *Routledge Encyclopedia of Philosophy*. Oxford: Routledge, 1998. Nine, p. 8600. doi: 10.4324/9780415249126-Q106-1. Cited in page 113.

GREENEMEIER, L. 20 years after *Deep Blue*: how AI has advanced since conquering chess. *Scientific American*, n. 2 June, 2017. Available at: <http://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/>. Cited in page 276.

GUNDERSON, K. Descartes, La Mettrie, language, and machines. *Philosophy*, v. 39, n. 149, p. 193–222, July 1964. Doi: 10.1017/S0031819100055595. Cited in page 160.

GUNDERSON, K. The imitation game. *Mind*, v. 73, n. 290, p. 234–45, 1964. doi: 10.1093/mind/LXXIII.290.234. Cited in page 122.

HACKING, I. Styles of scientific thinking or reasoning: A new analytical tool for historians and philosophers of the sciences. In: GAVROGLU, K.; CHRISTIANIDIS, J.; NICOLAIDIS, E. (Ed.). *Trends in the Historiography of Science*. Dordrecht: Springer, 1994, (Boston Studies in the Philosophy of Science). doi: 10.1007/978-94-017-3596-4_3. Cited 2 times in pages 32 and 33.

HARDY, G. H. Mathematical proof. *Mind*, v. 38, n. 149, p. 1–25, 1929. doi: 10.1093/mind/XXXVIII.149.1. Cited in page 221.

HARTREE, D. R. *Calculating instruments and machines*. Urbana: University of Illinois Press, 1949. Cited 5 times in pages 74, 143, 144, 145, and 172.

HAUGELAND, J. Understanding natural language. *Journal of Philosophy*, v. 76, n. 11, p. 619–32, 1979. doi: 10.2307/2025695. Cited in page 116.

HAYES, P.; FORD, K. Turing test considered harmful. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*. [S.l.: s.n.], 1995. p. 972–7. Cited 10 times in pages 15, 109, 111, 112, 116, 133, 134, 176, 177, and 190.

HICKEY, T. Accusations against Charles Chaplin for political and moral offenses. *Film comment*, v. 5, n. 4, p. 44–57, 1969. Available at: <http://www.jstor.org/stable/43754277>. Cited in page 48.

HILBERT, D. Mathematical problems. *Bulletin of the American Mathematical Society*, v. 8, n. 10, p. 437–80, 1902. doi: 10.1090/s0002-9904-1902-00923-3. Cited in page 221.

HILBERT, D.; ACKERMANN, W. *Principles of Mathematical Logic*. Chelsea, RI: American Mathematical Society, 1950 [1928]. English translation of *Grundzüge der Theoretischen Logik*, (Berlin: Springer, 1928). Cited in page 221.

HITCHCOCK, C. (Ed.). *Contemporary debates in philosophy of science*. Oxford: Blackwell Publishing, 2004. Cited in page 153.

HOARE, T. Edsger Wybe Dijkstra. *Physics Today*, v. 56, n. 3, p. 96, 2003. doi: 10.1063/1.1570789. Cited in page 116.

HODGES, A. *Alan Turing: The Enigma*. The centenary edition. Princeton: Princeton University Press, 2012 [1983]. Cited 18 times in pages 15, 23, 26, 28, 33, 43, 47, 48, 49, 51, 156, 168, 228, 233, 235, 236, 240, and 251.

HOLMES, R. *Prometheus*^{2}*: the two Shelleys and romantic science*. 2011. John Coffin Memorial Lecture delivered on 27 October 2011 at the University of London Institute of English Studies. Available at University of London School of Advanced Study channel: <http://www.youtube.com/watch?v=_ifHWz-Sljo>. Access on 13 Mar. 2020. Cited in page 25.

HUSBANDS, P.; HOLLAND, O. The Ratio Club: a hub of British cybernetics. In: HUSBANDS, P.; HOLLAND, O.; WHEELER, M. (Ed.). *The mechanical mind in history*. Cambridge, MA: MIT Press, 2008. cap. 6, p. 91–148. Cited 2 times in pages 51 and 240.

IRVINE, L. Foreword to the first edition. In: *Alan M. Turing: Centenary Edition*. Cambridge: Cambridge University Press, 2012 [1959]. p. xix–xxiv. Cited in page 27.

JEFFERSON, G. The mind of mechanical man. *British Medical Journal*, v. 1, n. 4616, p. 1105–10, 1949. Cited 22 times in pages 13, 20, 40, 47, 82, 92, 95, 118, 149, 150, 151, 154, 155, 156, 166, 167, 172, 246, 247, 248, 249, and 268.

JEFFERSON, G. René Descartes on the localisation of the soul. *Irish Journal of Medical Science*, n. 285, p. 691–706, September 1949. Cited in page 149.

JONES, A. Five 1951 BBC broadcasts on automatic calculating machines. *IEEE Annals of the History of Computing*, v. 26, n. 2, p. 3–15, 2004. doi: 10.1109/MAHC.2004.1299654. Cited in page 260.

JONES, A. Brains, tortoises, and octopuses: postwar interpretations of mechanical intelligence on the BBC. *Information & Culture*, v. 51, n. 1, p. 81–101, 2016. doi: 10.1353/lac.2016.0004. Cited in page 234.

KOYRÉ, A. Galilée et l'expérience de Pise. In: *Annales de l'Université de Paris*. [S.l.]: BnF, 1937. v. 12, n. 1 (Janvier-Février). Available at: <http://gallica.bnf.fr/ark:/12148/bpt6k938880>. Cited in page 185.

KOYRÉ, A. *Galileo Studies*. New Jersey: The Harvester Press, 1977 [1939]. Translated by J. Mepham from *Études galiléennes*, Paris: Hermann, 1939. Cited in page 185.

LAVINGTON, S. (Ed.). *Alan Turing and his contemporaries: building the world's first computers*. Swindon: British Conservation Society: the Chartered Institute for IT, 2012. Cited in page 256.

LEAVITT, D. *The man who knew too much: Alan Turing and the invention of the computer*. New York: W. W. Norton, 2006. (Great Discoveries). Cited 2 times in pages 23 and 50.

LIVIO, M. Did Galileo truly say, 'And yet it moves'? a modern detective story. *Scientific American*, 6 May 2020. Available at: <http://blogs.scientificamerican.com/observations/did-galileo-truly-say-and-yet-it-moves-a-modern-detective-story/>. Cited in page 55.

LUCAS, J. R. Minds, machines and Gödel. *Philosophy*, v. 36, n. 137, p. 112–27, 1961. doi: 10.1017/S0031819100057983. Cited in page 147.

MACH, E. On thought experiments. In: HIEBERT, E. N. (Ed.). *Knowledge and error: sketches on the psychology of enquiry*. Dordrecht-Holland: D. Reidel, 1976 [1897], (Vienna circle collection). cap. 11, p. 134–47. Translation of the 5th edition of *Erkenntnis und Irrturn* (Leipzig: Johann Ambrosius Barth, 1905), which included "Über Gedankenexperimente", in: *Zeitschrift für den physikalischen und chemischen Unterricht*, 10: 1–5., 1897. doi: 10.1007/978-94-010-1428-1. Cited 5 times in pages 152, 153, 176, 179, and 180.

MARCUS, G.; ROSSI, F.; VELOSO, M. Beyond the Turing test. *AI Magazine*, Association for the Advancement of Artificial Intelligence, v. 37, n. 1, p. 3–4, 2016. Special issue editorial. doi: 10.1609/aimag.v37i1.2650. Cited 4 times in pages 110, 115, 118, and 119.

MARTIN, C. D. ENIAC: press conference that shook the world. *IEEE Technology and Society Magazine*, v. 14, n. 4, p. 3–10, 1995. doi: 10.1109/44.476631. Cited in page 233.

MARTIN, E. *The calculating machines: their history and development*. Cambridge, MA: MIT Press, 1992 [1925]. English translation from Ernst Martin, *Die Rechenmaschinen und ihre Entwicklungsgeschichte*, (Germany: Pappenheim). Cited in page 221.

MAYS, W. Can machines think? *Philosophy*, v. 27, n. 101, p. 148–62, 1952. Cited 5 times in pages 27, 122, 128, 153, and 214.

MAYS, W. Turing and Polanyi on minds and machines. *Appraisal*, v. 3, n. 2, p. 55–62, 2000. Cited 3 times in pages 14, 251, and 254.

MAYS, W. My reply to Turing: fiftieth anniversary. *Journal of the British Society for Phenomenology*, v. 32, n. 1, p. 4–23, 2001. doi: 10.1080/00071773.2001.11007314. Cited 2 times in pages 27 and 256.

MCCARTHY, J.; SHANNON, C. Preface. In: SHANNON, C.; MCCARTHY, J. (Ed.). *Automata studies*. New Jersey: Princeton University Press, 1956. Cited 2 times in pages 117 and 118.

MCGILVRAY, J. Introduction to the third edition. In: (ED.), J. M.; (AUTHOR), N. C. (Ed.). *Cartesian linguistics: a chapter in the history of rationalist thought*. Cambridge: Cambridge University Press, 2009. p. 1–52. Cited in page 96.

MINSKY, M. *The society of mind*. New York: Simon & Schuster, 1985. Cited 2 times in pages 176 and 177.

MINSKY, M. *Marvin Minsky on AI: The Turing Test is a Joke!* 2013. Interview to the *Singularity Weblog*. Available at: <http://www.singularityweblog.com/marvin-minsky/>. Cf. from minute 23:35 to 24:45. Access on 9 Jul 2020. Cited in page 117.

MOOR, J. The status and future of the Turing test. *Minds and Machines*, v. 11, n. 1, p. 77–93, 2001. doi: 10.1023/A:1011218925467. Cited in page 133.

MOOR, J. H. An analysis of the Turing test. *Philosophical Studies*, v. 30, n. 4, p. 249–57, 1976. Doi: 10.1007/BF00372497. Cited 6 times in pages 62, 90, 98, 103, 123, and 124.

MOSCHOVAKIS, Y.; YATES, M. *In memoriam*: Robin Oliver Gandy 1919-1995. *The Bulletin of Symbolic Logic*, Association for Symbolic Logic, v. 2, n. 3, p. 367–70, 1996. Cited in page 106.

MOUNTBATTEN, L. The presidential address. *Journal of the British Institution of Radio Engineers*, v. 6, n. 6, p. 221–5, 1946. doi:10.1049/jbire.1946.0032. Cited 3 times in pages 26, 233, and 234.

MOUNTBATTEN, L. Address by the charter president – admiral of the fleet, the earl mountbatten of burma, k.g. *Proceedings of the Indian Division of the Institution of Electronic and Radio Engineers*, v. 2, n. 1, p. 9–14, 1964. Doi:10.1049/pidiere.1964.0003. Cited in page 235.

NAYLOR, R. H. Galileo and the problem of free fall. *British Journal for the History of Science*, v. 7, n. 2, p. 105–34, 1974. doi: 10.1017/S0007087400013108. Cited in page 186.

NEUBER, M. Feigl's 'scientific realism'. *Philosophy of Science*, v. 78, n. 1, p. 165–83, 2011. doi: 10.1086/658114. Cited in page 89.

NEUBER, M. Herbert Feigl. *Stanford Encyclopedia of Philosophy*, 2018 [2014]. Available at: <http://plato.stanford.edu/entries/feigl/#AnaMinBodPro>. First published 25 Apr. 2014. Substantive revision 12 Oct. 2018. Cited in page 95.

NEWMAN, M. H. A. *Letter to John von Neumann*. 1946. The Turing Archive for the History of Computing. Facsimile available at: <www.alanturing.net/newman_vonneumann_8feb46>. Cited in page 242.

NEWMAN, M. H. A. A note on electronic automatic computing machines. *British Medical Journal*, v. 1, n. 4616, p. 1133, June 25 1949. Cited 5 times in pages 17, 37, 137, 232, and 250.

NEWMAN, M. H. A. Alan Mathison Turing, 1912-1954. *Biographical memoirs of fellows of the Royal Society*, v. 1, n. November, 1955. Cited 8 times in pages 23, 32, 51, 53, 98, 220, 230, and 243.

NEWMAN, W. Max Newman: mathematician, codebreaker, and computer pioneer. In: COPELAND, B. J. (Ed.). *Colossus: the secrets of Bletchley Park's codebreaking computers*. Oxford: Oxford University Press, 2006. cap. 14, p. 176–88. Cited 2 times in pages 142 and 243.

NEWMAN, W. Alan Turing remembered: a unique firsthand account of formative experiences with Alan Turing. *Communications of the ACM*, v. 55, n. 12, p. 39–40, 2012. doi: 10.1145/2380656.2380682. Cited 4 times in pages 151, 248, 249, and 250.

NORTON, J. Are thought experiments just what you thought? *CanadianJournal of Philosophy*, v. 26, n. 3, p. 333–66, 1996. doi: 10.1080/00455091.1996.10717457. Cited 5 times in pages 181, 182, 183, 187, and 190.

NPL. *Minutes of the Executive Committee of the National Physical Laboratory for 23 Oct. 1945*. Teddington, 1945. Facsimile available at: <http://www.AlanTuring.net/npl_minutes_oct1945>. Acess on 27 July 2020. Cited in page 231.

NPL. *A.C.E. project – origin and early history*. Teddington, 1946. Public Record Office, Kew, Richmond, Surrey (document reference DSIR 10/385). Facsimile available at: <www.AlanTuring.net/ace_early_history>. Acess on 1 Dec. 2019. Cited in page 231.

OPPY, G.; DOWE, D. The Turing test. *Stanford Encyclopedia of Philosophy*, 2020 [2003]. Available at: <http://plato.stanford.edu/entries/turing-test/>. First published 9 Apr. 2003. Substantive revision 18 Aug. 2020. Cited in page 14.

PENROSE, R. *The Emperor's new mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford: Oxford University Press, 1989. Cited in page 147.

PETZOLD, C. *The annotated Turing: a guided tour through Alan Turing's historic paper on computability and the Turing machine*. Indianapolis: Wiley, 2008. Cited in page 220.

PICCININI, G. Turing's rules for the imitation game. *Minds and Machines*, v. 10, n. 4, p. 573–82, 2000. doi: 10.1023/A:1011246220923. Cited in page 133.

PINSKY, L. Do machines think about machines thinking. *Mind*, LX, n. 239, p. 397–398, 1951. Doi: 10.1093/mind/LX.239.397. Cited in page 122.

POLANYI, M. *Personal knowledge: towards a post-critical philosophy*. Second edition. Chicago: University of Chicago Press, 1974 [1958]. Cited 6 times in pages 96, 122, 147, 172, 253, and 256.

POPPER, K. *The logic of scientific discovery*. London: Routledge, 2002 [1959]. (Routledge Classics, vol. 56). English edition translated and extended from the 1935 German edition *Logik der Forschung: zur Erkenntnistheorie der Modernen Naturwissenschaft* (Wien: Springer-Verlag). Cited 4 times in pages 138, 163, 164, and 180.

PRICE, D. J. d. Automata and the origins of mechanism and mechanistic philosophy. *Technology and Culture*, v. 5, n. 1, p. 9–23, 1964. doi: 10.2307/3101119. Cited in page 42.

PRICE, J. V. *The ironic Hume*. Austin: University of Texas Press, 1965. Cited in page 31.

PROUDFOOT, D. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, v. 175, n. 5-6, p. 950–7, 2011. doi: 10.1016/j.artint.2011.01.006. Cited in page 176.

PROUDFOOT, D. Rethinking Turing's test. *The Journal of Philosophy*, v. 110, n. 7, p. 391–411, 2013. Cited 5 times in pages 64, 91, 98, 103, and 123.

PROUDFOOT, D. Mocking AI panic: Turing anticipated many of today's worries about super-smart machines threatening mankind. *IEEE Spectrum*, 2015. Cited 2 times in pages 28 and 38.

PROUDFOOT, D. The Turing test from every angle. In: COPELAND, B. J. et al. (Ed.). *The Turing Guide*. Oxford: Oxford University Press, 2017. cap. 27, p. 287–300. Cited 2 times in pages 65 and 91.

PROUDFOOT, D. Turing's concept of intelligence. In: COPELAND, B. J. et al. (Ed.). *The Turing Guide*. Oxford: Oxford University Press, 2017. cap. 28, p. 301–7. Cited 3 times in pages 65, 83, and 120.

PROUDFOOT, D.; COPELAND, J. Turing and the first electronic brains: what the papers said. In: SPREVAK, M.; COLOMBO, M. (Ed.). *The Routledge Handbook of the Computational Mind*. Oxford: Routledge, 2018, (Routledge Handbooks in Philosophy). cap. 2, p. 23–7. Cited 4 times in pages 26, 233, 234, and 235.

RAATIKAINEN, P. Gödel's incompleteness theorems. *Stanford Encyclopedia of Philosophy*, 2015. Available at: <http://plato.stanford.edu/entries/goedel-incompleteness/>. Accessed on 20 July 2020. Cited in page 222.

REINFELD, F. *Relax with Chess*. New York: Pitman, 2002 [1948]. Cited in page 240.

ROMERO, F. Cold War anti-communism and the impact of Communism on the West. In: NAIMARK, N.; PONS, S.; QUINN-JUDGE, S. (Ed.). *The Cambridge history of Communism*. Cambridge: Cambridge University Press, 2017. Vol. II: The socialist camp and world power 1941-1960s, cap. 12, p. 291–314. doi: 10.1017/9781316459850. Cited in page 48.

ROPE, C. Pioneer profiles: Douglas Hartree. *Computer Resurrection*, The Bulletin of the Computer Conservation Society, n. 49, Winter 2009/10 2010. Available at: <http://www.cs.man.ac.uk/CCS/res/res49.htm#d>. Access on 6 Nov. 2020. Cited in page 242.

RUSSELL, B. *A History of Western Philosophy*. New York: Simon & Schuster/Touchstone, 1972 [1945]. Cited 3 times in pages 130, 131, and 132.

RUSSELL, B. *Mysticism and logic and other essays*. London: George Allen & Unwin Ltd, 2008 [1910]. Available at Project Gutenberg's: <http://www.gutenberg.org/files/25447/25447-h/25447-h.htm>. Cited in page 84.

RUSSELL, B. *Our knowledge of the external world*. New York: Routledge, 2009 [1914]. (Routledge Classics). Cited in page 84.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence: a modern approach*. Third edition. New Jersey: Pearson, 2010 [1995]. (Prentice Hall Series in Artificial Intelligence). Cited in page 230.

SAMPLE, I.; HERN, A. Scientists dispute whether computer 'Eugene Goostman' passed Turing test. *The Guardian*, n. 9 jun., 2014. Available at: <http://www.theguardian.com/technology/2014/jun/09/scientists-disagree-over-whether-turing-test-has-been-passed>. Access on 20 Nov. 2020. Cited in page 119.

SAYGIN, A. P.; CICEKLI, I.; AKMAN, V. Turing test: 50 years later. *Minds and Machines*, v. 10, n. 4, p. 463–518, 2000. doi: 10.1023/A:1011288000451. Cited in page 123.

SCHURR, P. H. *So that was life: a biography of Sir Geoffrey Jefferson, master of the neurosciences and man of letters*. London: Royal Society of Medicine Press, 1997. Cited 2 times in pages 23 and 96.

SEARLE, J. Minds, brains, and programs. *Behavioral and Brain Sciences*, v. 3, n. 3, p. 417–24, 1980. doi: 10.1017/S0140525X00005756. Cited 2 times in pages 116 and 122.

SEGRE, M. Galileo, Viviani and the Tower of Pisa. *Studies in History and Philosophy of Science Part A*, v. 20, n. 4, p. 435–51, 1989. doi: 10.1016/0039-3681(89)90018-6. Cited 2 times in pages 183 and 185.

SHELLEY, M. *Frankenstein; or, the modern Prometheus*. Second edition. New York: W. W. Norton, 2012 [1818]. (Norton Critical Editions). Cited in page 25.

SHELLEY, P. B. *Shelley's "Prometheus Unbound": a variorum edition*. Seattle: University of Washington Press, 1959 [1818–1822]. (Ed.) Lawrence John Zillman. Cited in page 24.

SIMPSON, E. H. Introducing Banburismus. In: COPELAND, B. J. et al. (Ed.). *The Turing Guide*. Oxford: Oxford University Press, 2017. cap. 13, p. 129–42. Cited in page 230.

SLOWIK, E. Descartes' physics. *Stanford Encyclopedia of Philosophy*, 2017 [2005]. Available at <http://plato.stanford.edu/entries/descartes-physics/>. Cited in page 114.

SORENSEN, R. A. *Thought experiments*. Oxford: Oxford University Press, 1992. Cited in page 152.

STRAUSS, L. *Natural Right and History*. Reissue edition. Chicago: University of Chicago Press, 1999 [1953]. (Walgreen Foundation Lectures). Cited in page 48.

SUMNER, J. Defiance to compliance: visions of the computer in postwar Britain. *History and Technology*, v. 30, n. 4, p. 309–33, 2014. doi: 10.1080/07341512.2015.1008962. Cited in page 274.

SWINTON, J. *Alan Turing's Manchester*. Manchester: Infang Publishing, 2019. Cited 4 times in pages 23, 145, 250, and 251.

SYKES, C. *BBC Horizon: the strange life and death of Dr. Turing*. 1992. Documentary. Produced by Christopher Sykes, and edited by Jana Bennett. IMDb title 2373297, cf. <http://www.imdb.com/title/tt2373297/>. Available at: <http://www.youtube.com/watch?v=Z-sTs2o0VuY>. Access on 9 Jul 2020. Cited 3 times in pages 29, 72, and 231.

THURSTON, J. B. Devaluing the human brain. *The Saturday Review*, n. April 23, 1949. Cited 2 times in pages 150 and 247.

TIMES. Calculus to sonnet. *The London Times*, p. 4, 1949. (11 June). Cited 2 times in pages 151 and 249.

TIMES. No mind for mechanical man. *The London Times*, p. 2, 1949. (10 June). Cited 2 times in pages 150 and 249.

TIMES. Obituary for Dr A.M. Turing. *The London Times*, 16 June 1954. Cited in page 44.

TURING, A. *Letter to Ross Ashby*. 1946. W. Ross Ashby Digital Archive. Also at Woodger Papers (catalogue reference M11/99). Facsimile available at: <http://www.rossashby.info/letters/turing.html>. Access on 20 July 2020. Cited 3 times in pages 72, 232, and 237.

TURING, A. Letter from King's College, Cambridge. In: COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004 [*c.* 1940]. cap. Letters on logic to Max Newman, p. 214–6. Cited 4 times in pages 90, 167, 227, and 228.

TURING, A. Letter to Norman Routledge. In: HODGES, A. (Ed.). *Alan Turing: The Enigma*. Princeton: Princeton University Press, 2012 [*c.* early 1952]. p. xxviii. Preface to the centenary edition. Cited 2 times in pages 50 and 263.

TURING, A. et al. Rough draft of the Discussion on the Mind and the Computing Machine, held on Thursday, 27th October, 1949, in the Philosophy Seminar. *The Rutherford Journal*, v. 1, n. December, 2005 [1949]. Transcript edited by Jack Copeland of notes taken during the philosophy seminar co-chaired by Michael Polanyi and Dorothy Emmet at the University of Manchester on 27 October 1949, made available by Wolfe Mays. Available at: <http://rutherfordjournal.org/article010111.html>. Facsimile at: <http://www.alanturing.net/philosophy_seminar_oct1949/>. Access on 20 July 2020. Cited 15 times in pages 30, 57, 58, 80, 81, 97, 122, 145, 146, 147, 251, 252, 253, 254, and 255.

TURING, A. M. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42, n. 1, p. 230–65, 1936. doi: 10.1112/plms/s2-42.1.230. Cited 15 times in pages 13, 16, 18, 35, 39, 93, 137, 170, 216, 217, 218, 220, 223, 224, and 234.

TURING, A. M. Systems of logic based on ordinals. *Proceedings of the London Mathematical Society*, s2-45, n. 1, p. 161–228, 1939 [1938]. doi: 10.1112/plms/s2-45.1.161. Cited 5 times in pages 96, 224, 225, 226, and 227.

TURING, A. M. *Report on visit to U.S.A., January 1st - 20th, 1947*. Teddington, 1947. Public Record Office (document reference DSIR 10/385). Facsimile available at: <<http://www.alanturing.net/turing_usa_visit/>>. Access on 20 July 2020. Cited in page 241.

TURING, A. M. Computing machinery and intelligence. *Mind*, LIX, n. 236, p. 433–60, 1950. doi: 10.1093/mind/LIX.236.433. Cited 73 times in pages 13, 15, 17, 18, 26, 28, 30, 32, 33, 34, 35, 36, 37, 44, 45, 46, 59, 61, 67, 70, 71, 74, 78, 80, 82, 90, 93, 94, 97, 107, 116, 122, 124, 125, 127, 128, 129, 130, 131, 133, 134, 135, 137, 139, 143, 144, 147, 150, 153, 156, 157, 158, 159, 160, 161, 165, 166, 167, 168, 169, 170, 173, 174, 175, 191, 224, 237, 241, 246, 249, 256, 257, and 259.

TURING, A. M. *The programmers' handbook for the Manchester electronic computer*. 1951. Turing Digital Archive. Facsimile available at: <http://www.turingarchive.org/browse.php/b/32>. Cited in page 244.

TURING, A. M. Lecture on the Automatic Computing Engine. In: COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004 [1947]. p. 378–94. Chapter 9. Facsimile available at <http://www.turingarchive.org/browse.php/B/1>. Cited 13 times in pages 35, 44, 73, 128, 141, 144, 145, 147, 170, 214, 237, 238, and 239.

TURING, A. M. Intelligent machinery. In: COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004 [1948]. p. 410–32. Chapter 10. Facsimile available at: <http://www.turingarchive.org/browse.php/C/11>. Cited 25 times in pages 26, 36, 41, 45, 46, 53, 54, 65, 66, 73, 90, 93, 121, 128, 135, 140, 141, 148, 169, 174, 214, 239, 244, 245, and 246.

TURING, A. M. Can digital computers think? In: COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004 [1951]. p. 482–6. Chapter 13. Facsimile available at: <http://www.turingarchive.org/browse.php/b/5>. Access on 20 July 2020. Cited 16 times in pages 38, 41, 43, 69, 70, 71, 75, 76, 77, 78, 79, 91, 143, 260, 261, and 262.

TURING, A. M. Chess. In: COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004 [1953]. p. 569–75. Chapter 16. Facsimile available at: <http://www.turingarchive.org/browse.php/B/7>. Access on 20 July 2020. Cited 3 times in pages 274, 275, and 276.

TURING, A. M. Intelligent machinery, a heretical theory. In: COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004 [*c.* 1951]. p. 472–5. Chapter 12. Facsimile available at: <http://www.turingarchive.org/browse.php/B/20>. Access on 20 July 2020. Cited 12 times in pages 37, 41, 50, 74, 75, 120, 131, 175, 214, 258, 259, and 260.

TURING, A. M. Proposed electronic calculator. In: COPELAND, B. J. (Ed.). *Alan Turing's Automatic Computing Engine: the master codebreaker's struggle to build the modern computer*. Oxford: Oxford University Press, 2005 [1945]. p. 369–454.

doi: 10.1093/acprof:oso/9780198565932.003.0021. Chapter 20. Facsimile available at <http://www.alanturing.net/proposed_electronic_calculator/>. Access on 30 Sep. 2019. Cited 2 times in pages 71 and 232.

TURING, A. M. et al. Can automatic calculating machines be said to think? In: COPELAND, B. J. (Ed.). *The essential Turing: the ideas that gave birth to the computer age*. Oxford: Oxford University Press, 2004 [1952]. p. 494–506. Chapter 14. Facsimile available at: http://www.turingarchive.org/browse.php/B/6. Cited 19 times in pages 62, 67, 68, 70, 75, 76, 122, 129, 136, 153, 157, 263, 264, 265, 266, 267, 268, 269, and 270.

TURING, D. *Prof: Alan Turing Decoded*. London: Pitkin Publishing, 2015. Cited in page 23.

TURING, S. *Alan M. Turing: Centenary Edition*. Cambridge: Cambridge University Press, 2012 [1959]. Cited 13 times in pages 18, 23, 24, 27, 29, 30, 34, 39, 40, 42, 48, 52, and 230.

VIVIANI, V. Life of Galileo. In: FAVARO, A. (Ed.). *Le opere di Galileo Galilei, 20 vols*. Florence: Edzione Nazionale, 1968. v. 19. Cited 2 times in pages 184 and 185.

WALSHE, F. M. R. Geoffrey Jefferson, 1886-1961. *Biographical memoirs of fellows of the Royal Society*, v. 7, n. November, 1961. doi: 10.1098/rsbm.1961.0010. Cited 2 times in pages 149 and 246.

WEIZENBAUM, J. ELIZA: a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, v. 9, n. 1, p. 36–45, 1966. doi: 10.1145/365153.365168. Cited 2 times in pages 112 and 175.

WERRETT, S. Wonders never cease: Descartes's "météores" and the rainbow fountain. *The British Journal for the History of Science*, v. 34, n. 2, p. 129–47, 2001. Cited 2 times in pages 155 and 160.

WIENER, N. *Cybernetics: Or Control and Communication in the Animal and the Machine*. Second edition. Cambridge, MA: MIT Press, 1965 [1948]. Cited 7 times in pages 34, 40, 145, 150, 239, 240, and 247.

WIGNER, E. P.; HODGKIN, R. A. Michael Polanyi, 12 march 1891 - 22 february 1976. *Biographical memoirs of fellows of the Royal Society*, v. 23, p. 412–48, 1977. doi: 10.1098/rsbm.1977.0016. Cited in page 145.

WILKES, K. V. *Real people: personal identity without thought experiments*. Oxford: Oxford University Press, 1988. Cited in page 113.

WILLIAMS, F. C.; KILBURN, T. Electronic digital computers. *Nature*, v. 162, n. 4117, p. 487, 1948. 25 September. doi:http://doi.org/10.1038/162487a0. Cited 2 times in pages 13 and 242.

Appendix

APPENDIX A – Machine intelligence in Turing's thought (1936-1952)

From Turing's 1935-1936 invention of the abstract machine that he called "universal computing machine" up to his *c.* early1952 distressed syllogism and his *c.* late 1952 text on chess, over fifteen years have passed in the formation and development of an idea. What today is called "artificial intelligence," Turing liked to call *machine intelligence*. Back then in the earliest reception of Turing's 1950 paper, it was viewed by some as a paradoxical concept. Wolfe Mays (1952, p. 149), for instance, went to the dictionary to promptly show that Turing had just instilled nonsense. Thus read Mays from entry "machine" in the Oxford English Dictionary as of 1952:

> [A] combination of parts moving mechanically as contrasted with a being having life, consciousness and will. Hence applied to a person who acts merely from habit or obedience to a rule, without intelligence, or to one whose actions have the undeviating precision and uniformity of a machine.
> (MAYS, 1952, p. 149, as retrieved from the O.E.D. as of 1952)

And yet Turing did bring together "machine" and "intelligence" against all common sense. He wanted to challenge the conventional wisdom caught in common phrases such as "acting like a machine" (2004 [1947], p. 393), "purely mechanical behaviour" (2004 [1948], p. 410) or "you cannot make a machine to think for you" (2004 [*c.* 1951], p. 472), indeed. Right or wrong, in this chapter we shall see that Turing knew all too well what he was doing.

A.1 Problem and chapter structure

I will present a chronology of the idea of machine intelligence in Turing's thought according to the core sources, primary and secondary. As a result of the present chapter, we will have a chronological index (1936-1952) of Turing's views on machine intelligence (sometimes also in dialogue with others) to serve as reference to the chapters in the core part of this dissertation. My intention is to present a chronology of Turing's own views, not public intellectual history. Accordingly, I shall date Turing's papers, reports, lectures, letters and communications in general to the earliest time of production as indicated by the relevant primary or secondary sources.

In rendering this chronology I mostly rely on Jack Copeland's *The essential Turing* (2004). There one will find an anthology of Turing's core texts, all introduced from an integrated scientific, historical and philosophical point of view. So why take pains to produce this chronology myself? I shall now outline reasons why I felt that such a chronological index of Turing's concept of machine intelligence was in order. First, I needed to introduce to the reader in a self-contained way the Turing sources that are material to this dissertation. This is a convenience reason.

Beyond that, there are specific differences in my focus relative to Copeland's. His major task as I understand it was to put together the primary sources and provide essential intelligibility commentary also in light of plenty of secondary sources and the related literature. It was a huge historiographical effort that actually settled the basis for and made possible further works like mine. My attention in presenting the sources will be in turn *to follow the development of Turing's concept of machine intelligence distinctively and closely over time from source to source*. The specificity of this chronology shall prove itself fruitful. It lies in at least these five aspects: (i) the chronology will benefit from Copeland's findings and insights (I shall cite them specifically) so as to build upon them; (ii) I shall observe Turing's views closely in the context of his dialogue with core interlocutors such as Geoffrey Jefferson, Michael Polanyi, Max Newman and others; (ii) I will draw attention to specific aspects of Turing's views and those of his interlocutors from a point of view of the philosophy of science; (iv) the focus will be less on the analysis of Turing's views and more on their chronological development and his moves; and (v) I shall examine in detail specific elements of Turing's views in connection with my claims, as discussed shortly.

I propose the following periodizations of Turing's development of the concept of machine intelligence: (i) the period from Turing's 1936 paper up to his 1938 doctoral thesis at Princeton and moving back to Cambridge is mostly *foundational*; (ii) the one from Turing's wartime residence in Bletchley Park (1939) up to the delivery of his NPL 1948 report and settling in Manchester is mostly *experimental*; and (iii) the period from his interview to *The Times* in June 1949 on up to the BBC roundtable in January 1952 is mostly *dialogical*. The reader may keep these categories in mind as we go through the historical events. One could organize such a chronology in another way, say, in terms of Turing's mostly scientific (1936-1949) and mostly philosophical (1949-1952) years. My option, however, will draw attention to an important move from the point of view of the philosophy of science, which is *Turing's experimental turn* from mathematical logics to empirically-driven problem-solving during the war.

The chronology shall shed light on evidence for two claims that I intend to make:

- As a central and general claim, I hold that *Turing's views on machine intelligence were based on his science*, namely, his new mathematical science of computing and the mathematical techniques he developed to solve empirical problems during the war. So, in regard to Turing's statements when he spoke more distinctively philosophically in his dialogical years, I encourage the reader to keep track of the foundational and experimental elements that would, accordingly, underly and back up them. I will not structure this general claim analytically in this chapter. Recall that my goal here is to offer an index to serve to the chapters in the main part, where specific claims will be made. In those occasions I will refer to elements of Turing's thought that can be found indexed in here.

- As a specific claim, I hold that early on, since Turing's very first thoughts on machine intelligence during his wartime service through Turing's 1945, 1947 and 1948 NPL reports

and lectures up to the opening of the 1949 Manchester seminars, Turing still had the *game of chess* as his preferred choice of intellectual task to illustrate and test for machine intelligence. From his 1950 paper on up to the 1952 BBC roundtable and his syllogism in distress, Turing referred instead to *unrestricted conversation* as his chosen intellectual task. I would like to invite the reader to pay attention to this observation early on in this chapter. I shall develop it further and study its implications elsewhere (§3.5).

We shall now go through Turing's foundational (§A.2), experimental (§A.3) and dialogical (§A.4) years. At the end I will offer an analytical summary (§A.5) to sum up the key points in the development of Turing's concept of machine intelligence in connection with the two claims outlined above. I conclude with chapter acknowledgements (§A.6).

A.2 Foundational years (1936-1939): theorizing machines

There might exist a machine that could imitate some of the work done by the human mind. Turing conceived machines in terms of *(axiomatic) mathematical definition and proof*. For contrast, the reader may observe that earlier thinkers such as Blaise Pascal, Gottfried Leibniz and Charles Babbage all invented machines and sought support from craftsmen to build machines they described yet in terms of *material gears and wheels*. Their machines were physical. Their practical purpose was to automate calculations, so that they could improve business operations (Pascal), resolve cases in law, among other things (Leibniz), and tabulate "the constants of nature and art" at scale (Babbage). While Turing would later (from 1939 on and during the Second War World) share with them in part the purpose of automating calculations, his earlier purpose as of 1936 was essentially different, not practical but theoretical. Turing invented a machine to give mathematical form to the notion of mechanical process and prove a theorem on David Hilbert's *Entscheidungsproblem* ("decision problem"), which lived in the realm of pure mathematics. This problem was addressed in 1931 by Kurt Gödel who made significant progress and yet left it open. Turing gave a definitive answer to it. However, the technical results established by Gödel and himself still left some technical gaps with potential connections with mechanistic explanations of mathematical activity and the human mind. And Turing, indeed, tried yet to build upon Gödel's incompleteness theorems and push the field forward in his doctoral thesis at Princeton.

In his mostly foundational years, Turing set out to imitate the activity of the human mind and thus launched a disruptively new way to think about machines, as presented next.

A.2.1 Turing's abstract computing machines (May 1936)

On 28 May (1936), Turing's most important work was received by the London Mathematical Society (p. 230). With this note Turing opened his description of "[c]omputing machines" (§1):

> [T]he computable numbers are those whose decimals are calculable by finite means. [...] For the present I shall only say that the justification lies in the fact that the human memory is necessarily limited. (TURING, 1936, p. 231)

What Turing called "finite means" the reader may understand as systematic method or mechanical process (or, what today we call "algorithm"). The computable numbers are a subset of the real numbers. Their "decimal [fraction]," Turing wrote, "can be written down by a machine" (p. 230). That is, they can be calculated (we can say constructed) by a machine. A transcendental number such as π, for instance, is computable in the sense of Turing because there is a mechanical process that can be followed by a machine to construct it up to any desired decimal place. Over his assumption on human memory — it is "necessarily limited" —, Turing imagined a machine that would have a scanner or head endowed with some limited (main) "memory," consisting of a finite number of possible "states of mind." Supplied with a "tape," the machine would be able to receive input information from the world, keep track of erasable pieces of information produced during its operation, and record the new ones produced as its output. So the tape will serve also as a persistent memory. The head would go through the tape by observing symbols in its cells, one at a time. Now, the resemblance with the mental work of a human clerk working with pencil and paper is not inadvertent. I shall quote Turing to introduce the original language he conveyed for thinking about machines, before we are ready to see an example. Turing wrote:

> We may compare a man in the process of computing a real number to a machine which is only capable of a finite number of conditions q_1, q_2, ..., q_R which will be called "m-configurations" ['m' for "machine,"]. The machine is supplied with a "tape" (the analogue of paper) running through it, and divided into sections (called "squares"') each capable of bearing a "symbol". The machine is supplied with a "tape" (the analogue of paper) running through it, and divided into sections (called "squares") each capable of bearing a "symbol".
> (TURING, 1936, p. 231)

Turing initially called it an a-machine ('a' for "automatic," as opposed to c-machine with 'c' for "choice"). This distinction refers to batch and interactive computing. (For a contemporary reference, the reader may think of, say, a program that runs uninterruptedly in the background of their personal computer as an a-machine, as opposed to, say, a web browser that runs in response to a URL typed or to a link clicked by the user as a c-machine.) Given his 1936 focus on the (batch) computation of real numbers, Turing informed (1936, p. 231) that he would deal with the former only. This meant no loss of generality to his machine model, which is expressive enough to be amenable for adaptations. (One such adaptation is the conception of the "universal" machine, which we will see shortly.) Soon after Turing's 1936 publication, logician Alonzo Church referred to Turing's a-machine and eternized it as "a Turing machine" (1937).

The Turing machine

Figure 4 shows a scheme of the operation of a Turing machine that has four m-configurations: $\mathfrak{b}$, $\mathfrak{c}$, $\mathfrak{e}$ and $\mathfrak{f}$. (The gothic German typeface are Turing's. Machine configuration $\mathfrak{b}$ can be thought of

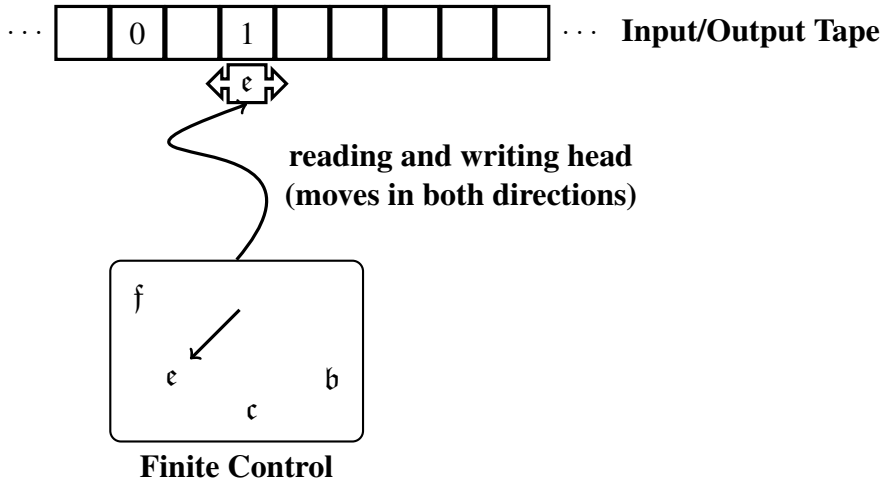

Figure 4 – The operation of a Turing machine that has four m-configurations, $\mathfrak{b}$, $\mathfrak{c}$, $\mathfrak{e}$ and $\mathfrak{f}$, and writes in the tape alternate figures of 0's and 1's separated by a blank square.[1]

as "begin." I shall walk the reader through the operation of this machine shortly.)

Turing introduced more language and details relative to the machine. About how to refer to the symbol in, say, cell or square r-th of the tape, he conveyed to use a capital S in gothic German font, $\mathfrak{S}(r)$. About how the machine would focus its attention and "remember," he wrote:

> At any moment there is just one square, say the r-th, bearing the symbol $\mathfrak{S}(r)$ which is "in the machine". We may call this square the "scanned square". The symbol on the scanned square may be called the "scanned symbol". The "scanned symbol" is the only one of which the machine is, so to speak, "directly aware". However, by altering its m-configuration the machine can effectively remember some of the symbols which it has "seen" (scanned) previously. (TURING, 1936, p. 231)

Finally, Turing considered whether a minimal set of conceptual resources was set for the machine behavior at each time to be fully determined. He introduced the basic notions about the manipulation of the tape:

> The possible behaviour of the machine at any moment is determined by the m-configuration q_n and the scanned symbol $\mathfrak{S}(r)$. [...] In some of the configurations in which the scanned square is blank (i.e. bears no symbol) the machine writes down a new symbol on the scanned square: in other configurations it erases the scanned symbol. The machine may also change the square which is being scanned, but only by shifting it one place to right or left. In addition to any of these operations the m-configuration may be changed. Some of the symbols written down will form the sequence of figures which is the decimal of the real number which is being computed. [...] (TURING, 1936, p. 231-2)

Turing concluded: "It is my contention that these operations include all those which are used in the computation of a number" (p. 232). We are now ready to review a Turing machine in its most

[1] This drawing is an adaptation of mine based on Sebastian Sardina's, whose code is available under LaTeX Project Public License at <http://texample.net/tikz/examples/turing-machine-2/>. Access on 20 July 2020.

Table 2 – Example of Turing machine (also referred to as "programme" or "instruction table").

Configuration		*Behavior*	
m-config.	*symbol*	*operations*	*final m-config.*
𝔟	None	$P0,\ R$	𝔠
𝔠	None	R	𝔢
𝔢	None	$P1,\ R$	𝔣
𝔣	None	R	𝔟

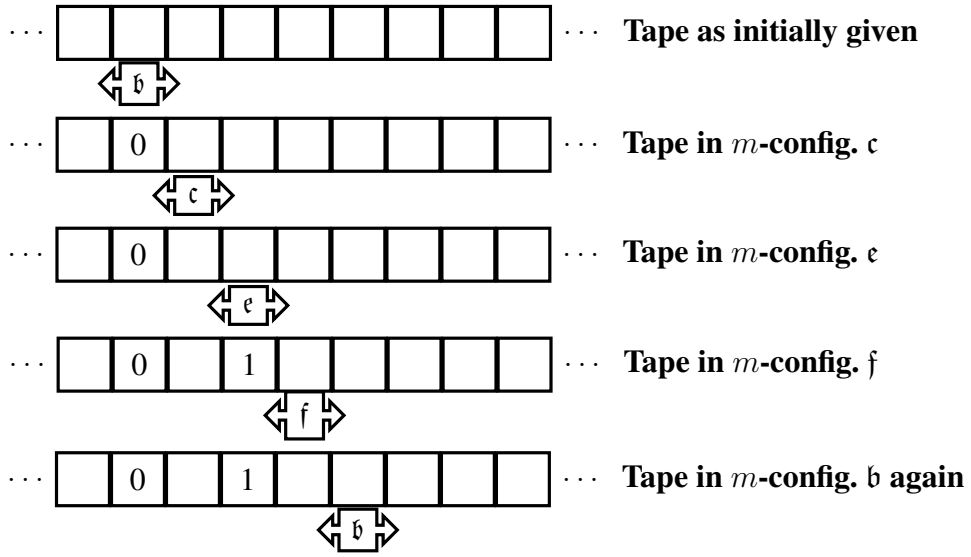

Figure 5 – Example of special-purpose Turing machine given in Turing's 1936 paper. This machine takes an empty tape (with "None" symbols) and writes in it alternate "figures" (0's and 1's) separated by a blank square. This sequence corresponds to the binary fraction (.01010101...) of the rational number 1/3 (whose decimal is .33333...).

basic structure and actions for illustrative purpose. I will reproduce the exact same names and symbols Turing used.

Although far from being an obvious abstraction, I shall say that a Turing machine is nothing but an "instruction table." (Yet Turing adopted this term only later, as pointed by Copeland in 2004, p. 30.) For its actual operation, the table must be supplied with the tape and the scanner (sensory-motor) mechanism. For example, recall the four-configuration machine shown in Figure 4 above. I shall now present in Table A.2.1 its instruction table and in Figure 5 five snapshots of its operation on the tape. In the beginning, the tape is empty (has only "None" symbols in it) and the machine is at the second square under m-configuration 𝔟 (see Figure 5, top). The machine then scans a "None" symbol, just as conditioned on the first entry of Table A.2.1, and its behavior will be fully determined by operations "$P0,\ R$" (print '0' and move right) and switch to m-configuration 𝔠). Now, following second entry of the table, the machine will scan the current square, which reads "None" indeed, and according to operation R will move right and switch to 𝔢. Then (third entry) after scanning another blank square, the machine will print '1' and move right again under m-configuration 𝔣. Now (fourth entry) it will move right and return

to m-configuration $\mathfrak{b}$. Note that as long as blank squares ("None") are scanned, the machine will not stop. It will just print 0's and 1's indefinitely. In fact, this machine was the first example given by Turing in (1936) and the sequence it computes corresponds to the rational number 1/3 written in binary (.01010101...), thus called by Turing a *computable sequence* or *number*.

The universal Turing machine

Now, from such an extraordinarily simple illustration of what his abstract machine can do Turing proceeded to show the mechanical construction of other numbers, with actual focus on transcendental numbers such as π. He developed several pieces of detail (further notation, conventions and codification schemes) so that the operations of the machine could be as conveniently encoded as possible. In particular, Turing observed that some computation tasks are common and may be subsidiary to different computable sequences). This was in fact the concept of subroutine, as identified by Copeland (2004, p. 12). Turing thus conveyed notation to abbreviate the expression of such common tables. He called them "skeleton tables" (p. 235) and (in order to refer to their use in other tables) "m-functions" (p. 236). Turing's intuition of this (in my words) compositionality principle was absolutely key towards the abstraction of the universal computing machine. Turing also conveyed an ingenious way to encode the table of any specific Turing machine (*e.g.*, Table A.2.1) as a single string of characters he called the "standard description" (of a given machine). Copeland noted that this corresponds to the process of compiling a computer program into machine code (2004, p. 12). Turing eventually reached the fundamental notion of the "universal computing machine." He made the announcement:

> It is possible to invent a single machine which can be used to compute any computable sequence. If this machine $\mathcal{U}$ is supplied with a tape on the beginning of which is written the S.D [standard description] of some computing machine $\mathcal{M}$, then $\mathcal{U}$ will compute the same sequence as $\mathcal{M}$. In this section I explain in outline the behaviour of the machine. (TURING, 1936, p. 241-2)

In other words, the universal machine can carry out *any* task that can be programmed into a Turing machine or instruction table. *Turing proved this result* by giving in section §7 of his 1936 paper a "detailed description" (the table of the behavior) of the universal machine. It is an existence proof by construction out of definitions. The table of the universal machine is composed of several m-functions or abbreviated tables he had introduced in previous sections. I shall now consider to have sufficiently presented its idea. In addition to Copeland's guide, the reader may find a dedicated and in fact very accessible presentation in Charles Petzold (2008). About the significance of the concept of the universal computing machine, Newman wrote: "[i]t is difficult to-day to realize how bold an innovation it was to introduce talk about paper tapes and patterns punched in them, into discussions of the foundations of mathematics" (1955, p. 256).

Leibniz, Hilbert, Gödel, and the purpose of Turing's "Computable numbers"

One might wonder why Turing came up with his new way to think about machines. The historical answer may trace back to Kurt Gödel and David Hilbert and if one likes may even be extended back to Gottfried Leibniz. The latter is said to have dreamt of a *universal characteristic* or language of symbols to express human thought and afford calculations over thought content (DAVIS, 2000). The rules of deduction would then be reduced to manipulations of such symbols and form a *calculus ratiocinator*, which Davis associates with today's symbolic logic. Leibniz must have considered that his calculator could be built into some actual machine. But he does not seem to have thought of any *abstract machine*. In fact, whenever the concept of machine is mentioned in connection with Leibniz, one means his *c.* 1672-1676 design of the "step reckoner" (also known as "Leibniz's wheel"), which would have extended Pascal's machine designs (MARTIN, 1992 [1925], p. 38-9). So Leibniz's designs of machines jumped from his abstract thoughts of a sort of symbolic logic straight to material gears and wheels.

In what follows I shall only briefly outline a formulation of connections between Hilbert's *Entscheidungsproblem* and the foundational results of Gödel and Turing's. My aim is to provide more of the historical background that may be relevant for the intelligibility of this dissertation.

In regard to Hilbert's program for mathematics, first there is Hilbert's outline of ten problems at the World Congress of Mathematics in Paris in 1900, which was published as an extended list of 23 problems in (1902). Then, there is his *Principles of mathematical logic* in collaboration with Wilhelm Ackermann in (1950 [1928]). There is where the *Entscheidungsproblem* or "decision problem" is posed. (Turing himself cited in his 1936 paper a German 1931 version of Hilbert and Ackermann's text. I am describing all this because often there is confusion in the literature. In 2013, p. 30, *e.g.*, David Anderson associated Hilbert's decision problem with his 1902 tenth problem — "[d]etermination of the solvability of a diophantine equation," Cf. HILBERT, 1902, p. 458 — of the list of 23 problems. But this is inaccurate.) Hilbert's decision problem asks for the existence of a general (systematic, effective, purely mechanical) process that takes a formula in the notation of a system L of symbolic logic as input and answers "yes" or "no" according to whether the formula is *universally valid* (1950 [1928]), that is, whether it is valid in every structure satisfying the axioms of L. If there were such a process, let me quote how G. H. Hardy described what the consequences for mathematics would be:

> Suppose, for example, that we could find a finite system of rules which enabled us to say whether any given formula was demonstrable or not. This system would embody a theorem of metamathematics. There is of course no such theorem, and this is very fortunate, since if there were we should have a mechanical set of rules for the solution of all mathematical problems, and our activities as mathematicians would come to an end. (HARDY, 1929)

Hardy's hunch was posed in early 1929, *n.b.*, before the advent of Gödel's 1931 incompleteness theorems. Martin Davis wrote that "Hilbert characterized" his decision problem "as the

fundamental problem of mathematical logic," because "it seemed clear to Hilbert" that with the solution of this problem "it should be possible at least in principle to settle all mathematical questions in a purely mechanical manner" (1965, p. 108). Now, the main difference between Hilbert and Leibniz's program was perhaps the scope of their intended calculation projects. While Hilbert's scope was the whole of mathematics, Leibniz wondered about some process like that to decide the truth or falsity of propositions out of just the whole of human thought.

The so-called *Gödel's incompleteness theorems* are two theorems Kurt Gödel showed in a (1965 [1931]) article. It will be important for us to have a rudimentary yet accurate understanding of them, since Turing presented results of his in connection with Gödel's, and also much discussion has been raised later relative to them in connection with Turing's views on machine intelligence (cf. §A.4.2). Let us first review how Gödel himself introduced his 1931 results:

> It is well known that the development of mathematics in the direction of greater precision has led to the formalization of extensive mathematical domains, in the sense that proofs can be carried out according to a few mechanical rules. (GÖDEL, 1965 [1931], p. 5)

He then mentioned Alfred Whitehead and Bertrand Russel's system of Principia Mathematica and the Zermelo-Fraenkel axiom system for set theory as the most extensive formal systems constructed back then, and completed:

> Both of these systems are so broad that all methods of proof used in mathematics today can be formalized in them, i.e., can be reduced to a few axioms and rules of inference. It is reasonable therefore to make the conjecture that these axioms and rules of inference are also sufficient to decide all mathematical questions which can be formally expressed in the given systems. In what follows it will be shown that this is not the case. [...] This situation does not depend upon the special nature of the constructed systems, but rather holds for a very wide class of formal systems [...]. (GÖDEL, 1965 [1931], p. 5-6)

Now, let us follow Panu Raatikainen's (2015) neat review of Gödel's results:

> *First incompleteness theorem.* Any consistent formal system F within which a certain amount of elementary arithmetic can be carried out is incomplete; i.e., there are statements of the language of F which can neither be proved nor disproved in F. (RAATIKAINEN, 2015)

According to Raatikainen, the above formulation accommodates an improvement due to J. B. Rosser in 1936. Note also the reference to "any consistent formal system." The difference with respect to Gödel's second theorem is that the latter concerns the limits of consistency proofs:

> *Second incompleteness theorem.* For any consistent system F [characterized as in the first incompleteness theorem], the consistency of F cannot be proved in F itself. (RAATIKAINEN, 2015)

Gödel's theorem is sometimes used to refer to the conjunction of these two, but may refer to either — usually the first — separately. Now, *a common misunderstanding*, Raatikainen added, *is to interpret Gödel's first theorem as showing that there are truths that cannot be proved*. This is incorrect, for the incompleteness theorem does not deal with provability in any absolute sense, but only concerns derivability in some particular formal system or another. In fact, somewhat vague references to Gödel's results gave rise later to varied forms of *Gödelian argument*. This refers in general to a view of Gödel's theorem, as lending itself to the philosophical implication that truth is superior to provability, or that the human mind is superior to machines. I will briefly present later (§A.2.2) Gödel's move that generated his two theorems, the so-called "Gödelization procedure."

Let us now shift to Turing and how he related Hilbert's decision problem and Gödel's theorems with his own results. Turing thus wrote:

> It should perhaps be remarked that what I shall prove is quite different from the well-known results of Gödel. Gödel has shown that (in the formalism of Principia Mathematica) there are propositions $\mathfrak{U}$ such that neither $\mathfrak{U}$ nor – $\mathfrak{U}$ is provable. [This is, *n.b.*, the *first incompleteness theorem*.] As a consequence of this, it is shown that no proof of consistency of Principia Mathematica (or of [the functional calculus] **K**) can be given within that formalism. [This is the *second incompleteness theorem*.] (TURING, 1936, p. 259)

Turing immediately proceeded to contrast Gödel's theorem(s) with his 1936 work:

> On the other hand, I shall show that there is no general method which tells whether *a given formula* $\mathfrak{U}$ is provable in **K**, or, what comes to the same, whether the system consisting of **K** with – $\mathfrak{U}$ adjoined as an extra axiom is consistent. (TURING, 1936, p. 259, emphasis added)

Now, in connection with what I have just emphasized in the passage above, let me make this difference that Turing pointed out straight. (I shall use Turing's notation and terms.) We shall first observe that Gödel's first incompleteness theorem is an existence proof. What he showed is that there will always be propositions $\mathfrak{U}$ (but which ones exactly?) that *cannot be* proved in Principia Mathematica. Other propositions *can be* proved in the system (but, again, which ones exactly?). Note that this does not solve Hilbert's *Entscheidungsproblem*, which asks for *a given formula* whether or not it is universally valid (meaning "provable") in the system. In fact, by means of his abstract machines and their properties as conceived in (1936), Turing solved Hilbert's decision problem himself. I am not in a position to go through it here but the answer is no — there is not a general process that could for such a given formula to answer whether or not it is universally valid or provable in the system it belongs. Turing continued from the above and completed:

> If the negation of what Gödel has shown had been proved, i.e. if, for each $\mathfrak{U}$, either $\mathfrak{U}$ or – $\mathfrak{U}$ is provable, then we should have an immediate solution of the Entscheidungsproblem. For we can invent a machine $\mathfrak{K}$ which will prove consecutively all provable formulae. Sooner or later $\mathfrak{K}$ will reach either $\mathfrak{U}$ or

> – 𝔘. If it reaches 𝔘, then we know that 𝔘 is provable. If it reaches – 𝔘, then, since **K** is consistent (Hilbert and Ackermann, p. 65), we know that 𝔘 is not provable. (TURING, 1936, p. 259)

That is, Turing meant, if Gödel had shown that all propositions 𝔘 were provable in the system, then he would have rendered the *Entscheidungsproblem* trivial — as a matter of course, we would know that each formula 𝔘 is provable in the system indeed. Another important fact to be noted about Gödel's theorem(s) is that, as we have seen by Raatikainen's remark above, it only concerns derivability in *some* particular formal system or another. In his 1938 doctoral thesis, Turing went further to study what would happen if multiple systems were combined.

A.2.2 Turing's doctoral thesis at Princeton (May 1938)

On 7 May 1938, according to Copeland (2004, p. 125), Turing's doctoral thesis *Systems of logic based on ordinals* was accepted at Princeton University under supervision of Alonzo Church. It was published one year later (1939 [1938]) by the London Mathematical Society. It is a lesser-known work of Turing's, yet an important one to shed light on his views on machine intelligence. It extended Turing's contributions in connection with Gödel's theorems (§A.2.1), with non-obvious implications for the question whether machines can think. Turing's thesis will be important in the context of his 1949 debate with Michael Polanyi (§A.4.2), and the discussion related to the so-called "mathematical objection" (1950) to machine thinking. In this exposition of Turing's thesis I shall rely in large part on Copeland's introduction (2004, Chapters 3, 4).

The purpose of ordinal logics

Most importantly for this chronology, Turing wrote in his (1939 [1938]) thesis:

> Mathematical reasoning may be regarded rather schematically as the exercise of a combination of two faculties [Turing's note: we are leaving out of account that most important faculty which distinguishes topics of interest from others; in fact, we are regarding the function of the mathematician as simply to determine the truth or falsity of propositions], which we may call *intuition* and *ingenuity*. (TURING, 1939 [1938], p. 214, no emphasis added)

This noteworthy philosophical distinction is key in connection with Gödel's argument and was introduced in his §11, "The purpose of ordinal logics," where Turing explained the vision of his thesis. He proceeded to characterize both concepts. On "intuition," Turing developed further:

> The activity of the intuition consists in making spontaneous judgments which are *not* the result of *conscious* trains of reasoning. These judgments are often but by no means invariably correct (leaving aside the question what is meant by 'correct'). Often it is possible to find some other way of verifying the correctness of an intuitive judgment. We may, for instance, judge that all positive integers are uniquely factorizable into primes; a detailed mathematical argument leads to the same result. This argument will also involve intuitive judgments, but they

> will be less open to criticism than the original judgment about factorization. I shall not attempt to explain this idea of 'intuition' any more explicitly.
> (TURING, 1939 [1938], p. 214-5, emphasis added)

Turing referred to consciousness as a boundary concept to dissociate intuition or "spontaneous judgements" from. For Turing, I interpret, intuition can be seen as a negative of consciousness. Also, Turing held, an intuitive judgement can be shown to be "correct" by means of an argument. We shall have a better sense of his notion of intuition in contrast with his concept of ingenuity:

> The exercise of ingenuity in mathematics consists in aiding the intuition through suitable arrangements of propositions, and perhaps geometrical figures or drawings. It is intended that when these are really well arranged the validity of the intuitive steps which are required cannot seriously be doubted.
> (TURING, 1939 [1938], p. 215)

For Turing, formal logics provided a best scenario to ground his notion of ingenuity. So he added:

> When working with a formal logic, the idea of ingenuity takes a more definite shape. In general a formal logic will be framed so as to admit a considerable variety of possible steps in any stage in a proof. Ingenuity will then determine which steps are the more profitable for the purpose of proving a particular proposition. (TURING, 1939 [1938], p. 215)

Altogether, I take, Turing thought that ingenuity involves elements of, say, selection, ordering, calculation and test in support of intuition. Unlike intuition, in my interpretation of Turing's reasoning, the exercise of ingenuity as existed in mathematics is typically conscious. Nonetheless, if intuition, on the one hand, may not be replaced (Gödel's incompleteness result), *ingenuity*, on the other hand, *can*. As pointed out by Copeland (2004, p. 136), *ingenuity can arguably be replaced by a suitably programmed Turing machine*. In fact, the ingenuity of a mathematician in sorting out lower level intuitions to compose the proof for a higher level proposition — *e.g.*, "all positive integers are uniquely factorizable into primes" — can be replaced by a mechanical process (an algorithm) that, given these intuitions and basic rules of inference as input, will deduce it all. Overall, Turing interpreted:

> In pre-Gödel times it was thought by some that it would probably be possible to carry this [Hilbert's] programme to such a point that all the intuitive judgments of mathematics could be replaced by a finite number of these rules. [...] We have been trying to see how far it is possible to eliminate intuition, and leave only ingenuity. We do not mind how much ingenuity is required, and therefore assume it to be available in unlimited supply. In our metamathematical discussions we actually express this assumption rather differently. We are always able to obtain from the rules of a formal logic a method of enumerating the propositions proved by its means. We then imagine that all proofs take the form of a search through this enumeration for the theorem for which a proof is desired. In this way ingenuity is replaced by patience. In these heuristic discussions [Ed.: Turing's own ongoing discussions through his 1938 thesis], however, it is better not to make this reduction. (TURING, 1939 [1938], p. 215)

One may note that Turing did not want to jump to conclusions in the "metamathematical discussions" related to Gödel's argument. Now we may be in a better position to appreciate what Turing ventured to accomplish in his thesis in connection with Gödel's incompleteness theorem.

Ordinal logics and Gödel's incompleteness theorem

First I shall briefly review what has been referred to in the literature as "the Gödelization procedure" (*e.g.*, Cf. DAVIS, 1993, p. 611). I will follow (and slightly adapt) Copeland's (2004, p. 138-9) exposition. Let S be the set of all arithmetical truths. We say that a formal-logic system L is complete with respect to S if every formula ℓ in S is provable in L. Now, as we have seen (§A.2.1), by Gödel's incompleteness theorem (1965 [1931]) there exist some arithmetical truths that cannot be proved in the formal system L. To prove it, Gödel found a way to construct an arithmetical formula ℓ' such that, by way of its own construction, it cannot be provable in the system and yet it is true. This shows L to be incomplete. Can we extend it to an L' that *completes* L with respect to the set S of all arithmetical truths by the addition of ℓ' as a new (ad-hoc) axiom that is then (trivially) provable in L'? No, because the Gödelization procedure can be applied recursively to instantiate now an ℓ'' that is true yet unprovable in L'. And this can be repeated *ad infinitum* to each new formal system L^i to which a missing formula is added.

Now, in his (1939 [1938]) thesis Turing set out to work around that in order to confine the significance of Gödel's theorems, or in other words, to establish the strictest boundaries of the place of intuition in mathematics. He introduced the concept of ordinal logics:

> The well-known theorem of Gödel [...] shows that every system of logic is in a certain sense incomplete, but at the same time it indicates means whereby from a system L of logic a more complete system L' may be obtained. By repeating the process we get a sequence $L, L_1 = L', L_2 = L'_1, ...$ each more complete than the preceding. A logic L_ω may then be constructed in which the provable theorems are the totality of. theorems provable with the help of the logics $L, L_1 = L_2, ...$ We may then form $L_{2\omega}$ related to L_ω in the same way as L_ω was related to L. Proceeding in this way we can associate a system of logic with any constructive ordinal. It may be asked whether a sequence of logics of this kind is complete in the sense that to any problem A there corresponds an ordinal α such that A is solvable by means of the logic L_a.
> (TURING, 1939 [1938], p. 161-2)

Turing's assumption was that such an ordinal logic would enable the distinction of the provable and the unprovable formulae of in each system within the larger context of the ordinal system. It would thus afford a formal distinction between intuitive steps and mechanical steps in the proof of number-theoretic theorems. Turing further wrote:

> We might hope to obtain some intellectually satisfying system of logical inference (for the proof of number-theoretic theorems) with some ordinal logic. Godel's theorem shows that such a system cannot be wholly mechanical; but with a complete ordinal logic we should be able to confine the non-mechanical steps entirely to verifications that particular formulae are ordinal formulae.

> We might also expect to obtain an interesting classification of number-theoretic theorems according to "depth". A theorem which required an ordinal α to prove it would be deeper than one which could be proved by the use of an ordinal β less than α. (TURING, 1939 [1938], p. 200-1)

Did Turing find interesting results? Yes, otherwise it would be unlikely to have created a new field of logics with an entry in an encyclopedia of philosophy. I will rather quote from Solomon Feferman, who wrote that entry (1998): "[f]or [the first ordinal logic to be considered in view of Gödel's results], Turing obtained a completeness result for the class of true statements of the form that all natural numbers have a given effectively decidable property." Nevertheless, the deeper meanings of that result in connection with the "metamathematical discussion" was not obvious. In his final note, Turing suggested that his ordinal-logic approach turns out to hit a tradeoff "between simplicity and comprehensiveness" (1939 [1938], p. 225). Feferman in turn assessed the limitations of Turing's thesis this way:

> [Turing] also showed that any ordinal logic (such as this) which is strictly increasing with increasing ordinal representation cannot have the property of invariance: in general, different representations of the same ordinal will have different sets of theorems attached to them. This makes the choice of representation a crucial one, and without a clear rationale as to how that is to be made, the notion of ordinal logic becomes problematic for its intended use. (FEFERMAN, 1998)

In fact, from the results of his thesis Turing himself drew only modest metamathematical implications, but enough to suggest a strategy to circunscribe intuition in mathematics. Turing discussed that in his correspondence with Max Newman, as presented next.

Ordinal logics and proof-finding machines

Also because it was written by using the exotic lambda-calculus notation designed by Alonzo Church, as of *c.* 1940 Newman was not yet familiar with Turing's thesis at Princeton. Turing wrote a letter to him and thereby provided significant further insight into its purpose:

> Ingenuity and Intuition. I think you take a much more radically Hilbertian attitude about mathematics than I do. You say 'If all this whole formal outfit is not about finding proofs which can be checked on a machine it's difficult to know what it is about.' When you say 'on a machine' do you have in mind that there is (or should be or could be, but has not been actually described anywhere) *some fixed machine* on which proofs are to be checked, and that the formal outfit is, as it were, about this machine. If you take this attitude (and it is this one that seems to me so extreme Hilbertian) there is little more to be said: we simply have to get used to the technique of this machine and resign ourselves to the fact that there are some problems to which we can never get the answer. On these lines my ordinal logics would make no sense.
> (TURING, 2004 [*c.* 1940], p. 215, emphasis added)

The reader may consider, for an example of "fixed machine" as Turing mentioned, Whitehead and Russell's Principia Mathematica system. As we know from Gödel's incompleteness theorem,

it will not be able to prove all true formulae of arithmetics. If that were the intention, Turing explained, then his ordinal logics would not be helpful. However, I interpret, by putting together and ordering various systems of logic, the formulae — unprovable in a system and provable in another system — would be all indexed in the system of the systems. Moreover, formal-logical systems and (Turing) machines could be thought of interchangeably, as he proceeded to explain:

> If you think of various machines I don't see your difficulty. One imagines different machines allowing different sets of proofs, and by choosing a suitable machine one can approximate 'truth' by 'provability' better than with a less suitable machine, and can in a sense approximate it as well as you please. The choice of a proof checking machine involves intuition, which is *interchangeable* with the intuition required for finding an Ω if one has an ordinal logic Λ, or as a third alternative one may go straight for the proof and this again requires intuition: or one may go for a proof finding machine.
> (TURING, 2004 [*c.* 1940], p. 215, emphasis added)

So Turing thought of his ordinal logic as a means to "approximate 'truth' by 'provability"' progressively, "as well as [one] please[s]." Note also that the "various machines," proof-checking, proof-finding and so on, would require only one actual machine to be realized — the universal machine. Indeed, I think it is implied in Turing's comment that the role of the universal machine is interchangeable with the role of his ordinal logic. And that is the key connection between Turing's 1936 and 1938 works.

Turing's experimental turn

Max Newman testified about Turing's interest in building a universal computing machine, and dated it to the time of Turing's 1936 paper. As Copeland related, Newman said to Christopher Evans in an interview that composes *The pioneers of computing: an oral history of computing* (available at the Science Museum in London): "Turing himself, right from the start, said it would be interesting to try and make such a machine." Also, Andrew Hodges collected from Newman's obituary "Dr. A. M. Turing" published by *The Times* on 16 June 1949:

> The description that he [Turing] then gave of a "universal" computing machine was entirely theoretical in purpose, but Turing's strong interest in all kinds of practical experiment made him even then interested in the possibility of actually constructing a machine on these lines.
> (HODGES, 2012 [1983], p. 109, note 2.38)

Indeed, *in the mostly experimental period of his intellectual life (1939-1949) Turing set out to build his universal computing machine, or,* as we shall see in detail in the next section, interchangeably in Turing's own words, *"to build a brain" and "to imitate a brain."* In regard to Turing's effort in his doctoral work, as we have seen, he tried to narrow down the significance of Gödel's 1931 incompleteness theorem(s) in connection with his views on machine intelligence.

There is a crucial point that I would like to draw attention to in the transition from Turing's foundational years to his experimental years. As of 1938, Turing was still tied to the

intellectual environment of pure mathematics in connection to Hilbert and Gödel's projects. His framework to think over the problem of imitating intuition and ingenuity in mathematics was that of formal logics. In that context, *certainty* cannot be sacrificed. The question of ensuring the *completeness* of a formal-logic system was only meaningful while preserving at the same time its *consistency* — otherwise the system's contradictions would entail any statement and thus render it as a whole trivial. Now, the same does not hold in the context of machine intelligence in general. In fact, *after the publication of his doctoral thesis and during his wartime service Turing learned that certainty or perfect accuracy is not required for intelligence*. The time pressure in Turing's mission in the military complex of Bletchley Park to break Nazi codes as fast as possible with less-than-perfect accuracy would be a key experience with deep implications on his concept of machine intelligence in general and to his reply to the mathematical objection to machine thinking based on Gödel's argument in particular.

A.3 Experimental years (1939-1949): building machines

According to the simplest view, Turing served his country in Wold War II by doing work on cryptography. No doubt Turing was recruited by the British Foreign Office with the mission of breaking the code of machine-cyphered German messages. Yet based on findings of Copeland's, there is another way to describe Turing's work in the war. What he did in this time was actually to pioneer *machine intelligence*, with code breaking as the specific application or intellectual task that was rather programmed for special-purpose machines to do. By doing that, in his wartime service (1939-1945) Turing developed his notion of machine intelligence significantly with respect to his foundational years. So much so that later, in the postwar period (1945-1949), he set out to further develop the machine-intelligence techniques that helped the Allies win the war, now to be implemented in a universal computing machine, seen like a brain. In particular, Turing learned in his experimental years a different path towards machine intelligence in comparison to his work on mathematical logics during his foundational years. In effect, while solving codebreaking by using heuristic techniques, Turing could observe that certainty is not required for the imitation of human intelligence, as presented next.

A.3.1 Turing's wartime service (Sep. 1939 – Jun. 1945)

Turing was recruited in the summer of 1938 to the Government Code and Cypher School (GC&CS), operated by the British Foreign Office (COPELAND, 2012, p. 35). He worked part-time from Cambridge on codes and ciphers since then. In September 1939, at the outbreak of the war, The GC&CS was moved to Bletchley Park, a secret military complex located in Buckinghamshire, England. Turing and his peers were based there to break encrypted radio messages from the Germans. These contained strategic military information, *e.g.*, positions of German submarines in the Atlantic and the date and time when they would attack British and American ships. As a means to carry out just that cryptography task, Copeland pointed out (2017b,

p. 266), Turing and colleagues developed an efficiently computable *heuristic technique* that he and colleagues called *banburismus*. Heuristics are built on the skillful use of shortcuts to prune the search space of a problem, and indeed turned to be a classical technique in modern artificial intelligence (Cf. RUSSELL; NORVIG, 2010 [1995], §4.2). About Turing, in connection with that, Max Newman wrote: "that combination of powerful mathematical analysis and intuitive short cuts that showed him at heart more of an applied than a pure mathematician" (1955, p. 255). Banburismus is named after Banbury, a market town where some special sheets they used to print the German messages came from (SIMPSON, 2017). Turing was familiar with the theoretical basis of heuristic problem-solving strategies, in particular the *minimax*, through John von Neumann's work on this topic (COPELAND, 2004, p. 563). The heuristic technique were programmed and run on special-purpose computers that they called *bombes* (2017a).

Now, chess was a popular game among code breakers. Copeland reports (2017b, p. 266, note 6) that he was told by Donald Michie and Irvin J. Good independently (both Turing's peers in the cryptography work at Bletchley Park) that Turing was concerned with the mechanization of thought processes involved in chess playing as early as of 1941. Michie told Copeland that he was a regular chess partner of Turing's, and that they used to talk about machine thinking when playing (COPELAND; PRINZ, 2017, p. 329). On Fridays, Michie said, after a week of work in an effort to break German figures, Turing and him, sometimes accompanied by others, went to a bar in the village of Wolverton (a few train stations from Bletchley Park). And in these meetings they would have discussed how to reproduce human thought processes on a universal Turing machine (*Ibid.*).It must have come naturally for Turing that, if they could make a machine to break codes using heuristics, maybe they could use similar techniques to make a machine to play chess. Michie also related that Turing would have circulated a draft on machine intelligence at Bletchley Park. This record, however, has not been preserved. It would be the oldest and most original in the history of machine intelligence (COPELAND, 2017b, p. 266).

Copeland's oral sources are consistent with what we know from Mrs. Sara Turing. She wrote (§1.5) that as early as of 1944 — that is, still during the war — Turing would have told her about his "about his plans for the construction of a universal computer and of the service such a machine might render to psychology in the study of the human brain" (2012 [1959], p. 92). In short, according to at least three independent secondary sources, *Turing was elaborating on machine intelligence still during the war, and possibly as early as 1941*. In particular, *he was thinking out how to program heuristics to make a machine break codes and play chess*.

Overall, about these 1938-1945 years Newman would recollect in his RS memoir:

> In 1938 Turing returned to Cambridge; in 1939 the war broke out. For the next six years he was fully occupied with his duties for the Foreign Office. These years were happy enough, perhaps the happiest of his life, with full scope for his inventiveness, a mild routine to shape the day, and a congenial set of fellow-workers. For his work for the Foreign Office he was awarded the O.B.E. [Order of the British Empire]. (NEWMAN, 1955, p. 254)

Regretfully, the state secrecy of Turing's work for the Foreign Office concealed a public exposition of his views, contributions and feats in relation to machine intelligence in that period.

A.3.2 Turing's NPL report on the ACE design (Dec. 1945)

After World War II, on the first of October 1945 (NPL, p. 6), Turing joined the National Physics Laboratory (henceforth NPL) to work on the construction of a machine that was called the Automatic Computing Engine (ACE) after Charles Babbage (1791-1871)'s project (COPELAND, 2004, p. 363). The official NPL documents have been assembled by Copeland (*Ibid.*) and include a chronology of events that led to Turing's recruitment (NPL, 1946). This chronology was put together by John R. Womersley, who recruited Turing for the NPL Maths Division. It is important historical material for my purpose here because, as already shown by Copeland, it connects Turing's 1936 paper with the British national project for building the ACE. Here is Womersley's chronology (At the end I added three events based on Copeland 2004, p. 364; 2012, p. 369):

> 1936-37 Publication of paper by A.M. Turing 'On Computable Numbers, with an Application to the Entscheidungsproblem' [...]
> 1937-38 Paper seen by J.R.W. [J. R. Womersley] and read. J.R.W. met C. L. Norfolk, a telephone engineer [...] and discussed with him the planning of a 'Turing machine' using automatic telephone equipment. Rough schematics prepared, and possibility of submitting a proposal to N.P.L. discussed. [...]
> Late 1943 J.R.W. first heard of [the] American machines.
> 1944 Sept. J.R.W. chosen for Maths. Division.
> 1945 Feb - May J.R.W. sent to the U.S.A. by Director. Sees Harvard machine and calls it 'Turing in hardware'. (Can be confirmed by reference to letters to wife during visit). J.R.W. sees ENIAC and is given information about EDVAC by Von Neumann and Goldstine.
> 1945 June J.R.W. meets Professor M. H. A. Newman. Tells Newman he wishes to meet Turing. Meets Turing same day and invites him home. J.R.W. shows Turing the first report on the EDVAC and persuades him to join N.P.L. staff, arranges interview and convinces Director and Secretary.
> 1945 Sept. Turing appointed; set to study reports on ENIAC and EDVAC. Turing decides that mechanisms proposed for EDVAC are appropriate to his ideas. J. R. W. begins negotiations with Post Office Engineering Research Stations.
> [1945 Oct. Turing officially starts.]
> 1945 Nov. Turing asked to write a report for the Executive Committee.
> [1945 Nov. - Dec. Turing writes the report.]
> [1946 Feb. Turing submits the report.]
> 1946 March Turing addresses Executive Committee. (NPL, 1946)

This NPL chronology establishes the connection between Turing's 1936 paper and Turing's mission at the NPL as recorded by his direct supervisor in the management staff. The ACE was meant to be the realization of a universal Turing machine. Now, what about Turing's own view of his job at the NPL? There is an interesting related testimony by Donald Bayley, who worked with Turing during the war between 1943 and 1946. When he joined the NPL, Bayley related, Turing said that he would build "a brain" (SYKES, 1992; COPELAND, 2004, minutes 25-27; p. 374). This is confirmed by a November 1946 letter from Turing to Ross Ashby, where Turing

wrote: "[i]n working on the ACE I am most interested in the possibility of producing models of the action of the brain than in practical applications to computing" (1946). Now, we shall observe that the NPL's institutional mission was all about practical applications of computing. So, one may foresee, the synergy between Turing and the NPL is bound not to last long (cf. §A.3.7).

A 1946 event that we will see next (§A.3.3)) would give public exposure for Turing's view of the ACE as (Britain's) "electronic brain." In any case, I take the analogy made by Turing in October 1945 between the ACE (machine) and the human brain as key evidence in connection with Womersley's chronology. More than having been hired to design a universal Turing machine for Britain (NPL staff's view), in his own view Turing joined the NPL to push forward his concept of machine intelligence. In short, by gathering these pieces, we can say that *Turing's concept of machine intelligence is historically tied through the ACE to his concept of the universal (Turing) machine in the history of his ideas*. The computer-mindbrain analogy is a basic assumption of the field of cybernetics that was emerging back then in the turn from the 1940's and the 1950's with Turing's contribution. And as we shall see, this analogy is the basis upon which the controversy with Jefferson was triggered (§A.4.1).

Back to the chronology of events at the NPL, we observe that in December 1945 Turing wrote up and submitted a report to the NPL executive committee with his proposed design for the ACE. It turns out that his *Proposed Electronic Calculator* report is the oldest preserved primary source where Turing mentioned the concept of machine intelligence. At some point in the report, Turing enumerated ten problems to which the ACE could be applied. It is in the last one, *problem 10*, where Turing related machine and intelligence towards a concept of machine intelligence:

> Given a position in chess the machine could be made to list all the 'winning combinations' to a depth of about three moves on either side. This is not unlike the previous problem [number 9], but raises the question 'Can the machine play chess?' It could be fairly easily be made to play a rather bad game. It would be bad because chess requires intelligence. We stated at the beginning of this section that the machine should be treated as entirely without intelligence. There are indications however that it is possible to make the machine to display intelligence at the risk of its making occasional serious mistakes. By following up this aspect the machine could probably be made to play very good chess. (TURING, 2005 [1945], p. 389)

So Turing, back then in late 1945, introduced a notion of machine intelligence and asked himself "Can the machine play chess?". Let us take a moment to recall Newman's distinction from the Introduction. Early at this point Turing is not asking whether machines can imitate "*all* kinds of thought, logical, poetical, reflective," but whether they can imitate "*anything* that can be called 'thought"' (1949). So as of 1945 Turing's concept of machine intelligence was still relatively modest, and would be further developed in depth and in breadth in Turing's following communications and writings (in 1947, 1948, 1950 and on) as we shall see below. Also in view of §3, there is a key point to observe here (in connection with what we have seen before (§A.3.1) and with what we will see later (§A.3.9). *Early from Turing's wartime service in 1941, and most*

clearly in late 1945 on, Turing worked with the game of chess as intellectual task and testbed for the development of his concept of machine intelligence. In particular, we find here a reference, even if perhaps a timid one, to the notion of allowing the machine to make mistakes. We shall see shortly that in 1947 Turing would relate it explicitly with the objection to machine thinking based on Gödel's argument.

In connection with Turing's work from late 1945 to 1946, there is a specific public event in Britain that attracted the attention of newspapers. It is the next event that I consider worth mentioning for the purpose of this chronology.

A.3.3 Louis Mountbatten's presidential address to the British Institution of Radio Engineers (Oct. 1946)

Since the end of the war, an arms race began for the construction of a modern computer. The NPL project to which Turing was recruited is one such initiative. Although much related information was held state secret, some was selected to be released to the press. This rendered some repercussion in the public domain, and a debate about what to expect from the new electronic computing machines. In (2018), Diane Proudfoot and Jack Copeland have brought forth a rich gathering of public events as covered in the British press (1938-1952) in this early history of modern computing. One of the most remarkable of those events was already noted by Andrew Hodges (2012 [1983], p. 347). It was a speech made by Admiral the Viscount Mountbatten of Burma (1946). Louis Mountbatten was an important British Royal Navy officer and statesman, an uncle of Prince Philip, Duke of Edinburgh. He had several prominent positions in the British Armed Forces. Now, it is a remarkable fact in the social history of modern computing that it was through a speech by this figure that Turing's concept of machine intelligence was brought forward in Britain's society and culture. That Mountbatten's historical speech is supplied with Turing's concept of machine intelligence, we know from evidence gathered by Hodges (2012 [1983], p. 347-9) and Proudfoot and Copeland (2018, p. 26-7). I take this as important because it means that the repercussion around Mountbatten's speech can be viewed actually as an anticipation of the reception of Turing's ideas in British society and culture. Let us first look at Mountbatten's discourse in some detail before we briefly review such evidence.

It was on 31 October 1946 that Mountbatten, then in the position of president of the British Institution of Radio Engineers, gave his elevated address. It came out with some delay in the wake of a major press conference held in Philadelphia in February the same year on the ENIAC (Cf. MARTIN, 1995), which was the largest related American project. He introduced the term "electronic brain," and outlined an overt perspective for the "thought-saving" technology of electronic computing machines (as opposed to "labour-saving"). These machines would perform (automate) increasingly more functions in industry. The analogy with the human mindbrain was quite explicit. The allusion to machine intelligence was unmistakable. Even a reference to science-fiction writer H. G. Wells was made. Thus Mountbatten said to the British engineers:

> [T]he stage is now set for the most Wellsian development of all: the Electronic Brain. [...] It is now considered possible to evolve an electronic brain which will perform functions analogous to those at present undertaken by the semi-automatic portion of the human brain. That is to say, it will receive information about the situation of the machinery under its control, and will provide an intelligent – I repeat, intelligent – link between that information and the action necessary to keep the machinery in general conformity with the overall directions given to it by man. [...] (MOUNTBATTEN, 1946, p. 223-4)

At some point, Mountbatten shifted his focus from industrial automation, so to speak, towards activities performed by human mathematicians that require "choice and discrimination" (p. 224). The resemblance with Turing's views is striking. Mountbatten even referred to the "hitherto human prerogatives" of choice and judgement, and to making a machine to play chess:

> [...M]achines now actually in use can exercise a degree of memory; and some are now being designed to exercise those hitherto human prerogatives of choice and judgment. One of them could even be made to play a rather mediocre game of chess! (MOUNTBATTEN, 1946, p. 224)

In fact, Mountbatten also alluded to "the abolition of routine mental work" and to us being enabled "to free the human mind for creative purposes" (p. 224). He said that "we are really facing a new revolution," now not an industrial one, "but a revolution of the mind" (*Ibid.*).

Now, let us briefly review the relation of Mountbatten's statements with Turing's ideas as we know from sources prior to the former's presidential address. First, the designation of an *electronic* computing machine as a *brain* had already been made by Turing in October 1945 (§A.3.2) and must have circulated within the NPL. Second, the observation of the intellectual activity of human mathematicians is also a core point that is originally from Turing's (1936) paper. Third, the idea of making a machine to able "to play a rather bad [mediocre] game" of chess is strongly related to Turing's point in his 1945 NPL report (§A.3.2).

Mountbatten used the American ENIAC project as example of electronic brain, and did not mention NPL's ACE. But his speech was disseminated the day next on the first of November 1946 by *The Times*. It had a lot of repercussion not only in the British press but also in international newspapers, often making their first page. I figure that all the secrecy about the piecemeal developments in the science and technology of electronic computing machines during the war must have contributed to shock the public. It was all of a sudden that the revelation of an electronic brain was broadcasted. For a survey on that, the reader may refer to Proudfoot and Copeland (2018). As it seems, Mountbatten's address forced the NPL to move on from total secrecy. Hodges reports (p. 348) that an official NPL press release on 6 November 1946, as reproduced on 8 November in *The Electrician*, introduced the ACE project and its roots in Turing's 1936 paper. According to Hodges, NPL's account was much more sober than Mountbatten's. Darwin, as reported by Allan Jones (2016, p. 84), also gave a talk on the BBC on the Home Service on 9 November to announce the project. Another week later, on 13 November

1946, NPL director Charles Darwin (homonymous and grandson of the great natural historian) wrote a letter to *The London Times*. In that letter, Proudfoot and Copeland (2018, p. 26-7) point out, Darwin explained that Mountbatten had been "fully informed" about the ACE, but at the NPL's request "did not mention it explicitly because it had not yet been made public." Also, said Darwin in the letter, Mountbatten described ACE's capacities "correctly in every respect." At this point, nonetheless, the repercussion in the press and its reader's correspondence was such that Darwin would also have to participate in a radio talk — "Did you hear that: an arithmetical robot" — in *The Listener* on 14 November 1946 about the ACE. In that talk, as also reported by Proudfoot and Copeland (p. 26-7), Darwin would have then emphasized that it was Turing's aim in his 1936 paper to discover the "ultimate limitations" of "a machine which would imitate the processes of thought." Furthermore, he said, wartime "developments" involving "electronic valves" were enabling Turing to show "how to make his idea come true" at the NPL. As Proudfoot and Copeland also found out (p. 26), in (1964, p. 14, only two decades after his 1946 address) Mountbatten said of this speech that he had "had the privilege to disclose for the first time some of the development work which had been undertaken during the war on the electronic computer."

In the wake of Mountbatten's address in November 1946, Turing did make an appearance. Hodges gathered (2012 [1983], p. 348-9) two interviews of Turing about the ACE. First, the *Daily Telegraph* reported on 7 November under headline "'ACE' will speed jet flying" an account based on interviews of John Womersley, Douglas Hartree and Turing at the NPL. Hartree, Fellow of the Royal Society since 1932, had recently, in 1946, took the post of Plummer professor of mathematical physics at the University of Cambridge. He would have said to the *Telegraph*: "[t]he implications of the machine are so vast that we cannot conceive how they will affect our civilisation." But Hartree meant practical applications of scientific computing. Turing, however, would have gone his own way: "Dr Turing, who conceived the idea of Ace, said that he foresaw the time, possibly in 30 years, when it would be as easy to ask the machine a question as to ask a man." The contrast between Hartree's view and Turing's view was marked. The former has also been attributed to have said that "the machine would always require a great deal of thought on the part of the operator." He also would have deprecated "any notion that Ace could ever be a complete substitute for the human brain." And this was the very occasion when Hartree would have implied a connection between views such as Turing's and Nazism. He would also have said to the *Telegraph*: "The fashion which sprung up in the last 20 years to decry human reason is a path which leads straight to Nazism."

That may have been the first newspaper interview of Turing. And he does not seem to have felt intimidated. The next day he received another reporter from lower-profile newspaper *Surrey Comet* willing to investigate on the "New NPL Wonder." In this other interview, Turing would have estimated his prediction to come true not in 30 but in 100 years. The story was published, Hodges relates,[2] on 9 November 1946 featuring "Dr. A. M. Turing, 34-year-old

[2] Hodges keeps facsimile of a clipping of the newspaper at his page: <http://www.turing.org.uk/scrapbook/ace.html>. Access on 20 July 2020.

mathematics expert, who is the pioneer of the scheme in this country." And it ran:

> Asked about Lord Louis Mountbatten's statement that it would be able to play an average game of chess, Dr Turing said that was looking far into the future. [...] The point was then put to him that chess and similar activities required judgment as well as memory, and Dr Turing agreed that that was a matter for the philosopher rather than the scientist. 'But,' he added, 'that is a question we may be able to settle experimentally in about 100 years time.'
> (HODGES, 2012 [1983], p. 349)

Turing's note about chess playing as an intellectual task, that its assessment is for the philosopher rather than the scientist, can be understood in light of what we have seen in section §1.4. His reference to experiment as the way to settle the question can be understood likewise.

A.3.4 Turing's letter to Ross Ashby (*c.* Nov. 1946)

Near enough to the events triggered by Mountbatten's address but not clearly related to it, Turing wrote a letter to psychiatrist and cybernetics pioneer W. Ross Ashby as mentioned (§A.3.2). In the Ashby archive it is guessed to date to 19 November 1946. Hodges dated it (2012 [1983], p. 360) to 20 November 1946, and Copeland is only certain (2004, p. 375, note 57) that it was written in between 1946 and 1947. The letter is, in any case, as I shall now suggest, a remarkable primary source about Turing's concept of machine intelligence as early as 1946-1947.

A first point that I would like to highlight in Turing's letter to Ashby is his introductory sentence in relation to Darwin, NPL's director: "Sir Charles Darwin has shown me your letter, and I am most interested to find that there is someone working along these lines." From that, I deduce that Turing must have been referring to some previous letter from Ashby *to Darwin or to the NPL*. Apparently at this point Turing and Ashby had never met. Now, since Darwin showed Ashby's letter to Turing, and being Ashby a psychiatrist and cybernetician, I take Turing's letter to Ashby as evidence that Darwin must have known about and may even have supported Turing's *electronic brain project* view back then, and *n.b.*, not only the *universal machine project* view.

In this letter, after confessing to Ashby about his intentions with the ACE at the NPL, namely, not "the practical applications to computing" but "producing models of the action of the brain" (§A.3.2), Turing proceeded to a discussion about the machine itself and how he was planning to study its possibilities. This is in fact the core theme of the letter and would indeed last for two paragraphs until its end. Essentially two topics are addressed, namely, (i) *a baseline approach for machine learning* and its limitations, and in connection with that, (ii) *the extension and limits of the machine as an analogue model of the brain*. So it compels us to date to late 1946 some of the key points of Turing's 1950 paper and his views on machine intelligence in general. This is the crucial importance of the letter, which is to my knowledge yet to be appreciated in the literature with regard to Turing's concept of machine intelligence. I shall then quote below large excerpts of the letter, marking Turing's move from topic (i) to topic (ii):

> The ACE will be used, as you suggest, in the first instance in an entirely disciplined manner, similar to the action of the lower centres [of the brain], although the reflexes will be extremely complicated. The disciplined action carries with it the disagreeable feature, which you mentioned, that it will be entirely uncritical when anything goes wrong. It will also be necessarily devoid of anything that could be called originality. (TURING, 1946)

So Turing acknowledged to Ashby that he was considering as well the idea of experimenting with the ACE, first of all, "in an entirely disciplined manner", similarly to "the lower centres" of the brain. This would be a specific educational approach, say, based on conditioned reflexes, that would bring with it the limitation of lending the machine incapable of what Turing would later in (1950), citing Hartree, call to "think for itself" (p. 450). At this point, however, he started a relatively long rebuttal to the idea that such limitation would be a necessary limitation of the ACE in general as a universal machine:

> There is, however, no reason why the machine should always be used in such manner: there is nothing in its construction which obliges us to do so. It would be quite possible for the machine to try out variations of behaviour and accept or reject them in the manner you describe and I have been hoping to make the machine do this. This is possible because, without altering the design of the machine itself, it can, in theory at any rate, be used as a model of any other machine, by making it remember a suitable set of instructions. (TURING, 1946)

Turing was, indeed, excited about experimenting with *the ACE as a digital computer*, which would not need hardware redesign in order to adapt its behavior. He was clearly trying to think out programming approaches to test the possibilities of the machine that would be the first realization of his abstract universal machine. He thus continued and finished along these lines:

> The ACE is, in fact, analogous to the 'universal machine' described in my paper on computable numbers. This theoretical possibility is attainable in practice, in all reasonable cases, at worst at the expense of operating slightly slower than a machine specially designed for the purpose in question. Thus, although the brain may in fact operate by changing its neuron circuits by the growth of axons and dendrites, we could nevertheless make a model, within the ACE, in which this possibility was allowed for, but in which the actual construction of the ACE did not alter, but only the remembered data, describing the mode of behaviour applicable at any time. [...] (TURING, 1946)

A.3.5 Turing's NPL lecture in London (Feb. 1947)

On 20 February 1947, Turing delivered a lecture on the ACE design for the London Mathematical Society (2004 [1947]). The lecture was one of a series of nine lectures given by Turing and his assistant Jim Wilkinson during the period December 1946 to February 1947. For a detailed account the reader may refer to (COPELAND et al., 2005, p. 459). My goal here is to emphasize, as of 1947, Turing's views on machine intelligence and on the ACE as a realization of Turing's 1936 concept of universal computing machine. About the latter, as pointed out by Copeland (2004, p. 16), early in his lecture Turing observed:

> [...D]igital computing machines such as the ACE [...] are in fact practical versions of the universal machine. There is a certain central pool of electronic equipment, and a large memory. When any particular problem has to be handled the appropriate instructions for the computing process involved are stored in the memory of the ACE and it is then 'set up' for carrying out that process. (TURING, 2004 [1947], p. 383)

Turing saved the last part of his 1947 lecture to disseminate his views on machine intelligence. We have seen before (§1.6) that Turing's plea for fair play for the machine was a high moment of that part of the lecture. He also gave on that occasion a preliminary rebuttal for what he named in 1950 the mathematical objection and Lady Lovelace's objection (cf. Figure 1). In connection with machine intelligence, *Turing addressed as a main topic of his 1947 lecture the storage capacity of the ACE as a universal Turing machine. For him, that would be a key property for the machine to be able to display some intelligence*. Given his analogy with the human brain, Turing used to just call it "memory" (and in fact this is how we call it today but back then he received antagonism for doing so and seems to have sometimes tried to abide by it). In any case, Turing's concern was with the minimal amount of storage units needed in order to make his initial experiments on machine intelligence to become feasible. At some point of his February 1947 talk, Turing said:

> I have spent a considerable time in this lecture on this question of memory, because I believe that the provision of proper storage is the key to the problem of the digital computer, and certainly if they are to be persuaded to show any sort of genuine intelligence much larger capacities than are yet available must be provided. (COPELAND, 2004, p. 383)

And, indeed, Turing's analogy with the human brain was workable. It was, *e.g.*, the source of his estimate about the minimal amount of storage needed for intelligence. Thus he reasoned:

> [...] The memory capacity of the human brain is probably of the order of ten thousand million [10^{10}] binary digits. But most of this is probably used in remembering visual impressions, and other comparatively wasteful ways. One might reasonably hope to be able to make some real progress with a few million [10^6] digits, especially if one confined one's investigations to some rather limited field such as the game of chess. (TURING, 2004 [1947], p. 393)

This passage is important for us to keep track of Turing's estimates about how far he thought he was to be able to run experiments on machine intelligence. But it is also a key source as of early 1947 for his rationale about chess playing as the intellectual task.

Turing also addressed the objection to machine thinking based on Gödel's argument:

> It has for instance been shown that with certain logical systems there can be no machine which will distinguish provable formulae of the system from unprovable, i.e. that there is no test that the machine can apply which will divide propositions *with certainty* into these two classes. Thus if a machine is made for this purpose it must in some cases fail to give an answer. On the other hand if a

> mathematician is confronted with such a problem he would search around a[nd] find new methods of proof, so that he ought eventually to be able to reach a decision about any given formula. [...Gödel's and related] theorems say nothing about how much intelligence may be displayed if a machine makes no pretence at infallibility. (TURING, 2004 [1947], p. 393, emphasis added)

Recall that this has been a main topic of his 1938 doctoral thesis at Princeton. As I have mentioned in the transition from his foundational to these experimental years, eventually Turing observed that the certainty required in a formal-logic system is not actually required for a machine to display intelligence. If in his 1945 NPL report Turing touched on this topic rather in passing when discussing the capability of the machine to play chess, in 1947 he addressed it directly.

Finally, in the closing of his lecture, Turing would again emphasize his choice of the game of chess for his experiments on machine intelligence:

> [...] the machine must be allowed to have contact with human beings in order that it may adapt itself to their standards. The game of chess may perhaps be rather suitable for this purpose, as the moves of the machine's opponent will automatically provide this contact. (TURING, 2004 [1947], p. 394)

In short, from Turing's 1947 comments on the game of chess, we observe: (i) being a limited (well-defined) field, the game of chess requires from the machine relatively low storage capacity (10^6 units) than that of the human brain (of the order of 10^{10} units); and (ii) to display intelligence the machine will have to adapt to (say, learn) human standards, and again the game of chess seems suitable, as the moves of the machine's opponent will automatically provide such contact (say, data for learning). In his next communication (2004 [1948]), as we will see shortly (§A.3.9), Turing would address in more detail and more comprehensively his rationale.

A.3.6 Meeting with Norbert Wiener (Spring 1947)

It was the spring of 1947. Norbert Wiener (1894-1964), then Professor at the Massachusetts Institute of Technology (MIT), was a leader in the field of study that he himself was about to coin "cybernetics, or control and communication in the animal and the machine." He stopped by England and met Turing. We know it from Wiener's *Cybernetics* (1965 [1948]), where Wiener acknowledged Turing highly and wrote about his trip through England:

> In the spring of 1947, I received an invitation to participate in a mathematical conference in Nancy on problems arising from harmonic analysis. I accepted and, on my voyage there and back, spent a total of three weeks in England, chiefly as a guest of my old friend Professor J. B. S. Haldane. I had an excellent chance to meet most of those doing work on ultra-rapid computing machines, especially at Manchester and at the National Physical Laboratories at Teddington, and above all to talk over the fundamental ideas of cybernetics with Mr. Turing at Teddington. (WIENER, 1965 [1948], p. 23)

At the time *Cybernetics* was published (22 October 1948), Wiener had already coined his favorite term (claimed to have been done in the summer of 1947, so after his stay in England; Cf. *Ibid.*,

p. 12) and seems to be doing his best to encompass within it everything related to the new science and technology of electronic computing machines. But in fact the British "cybernetics" [*sic*] hub of which Turing was member would later in July 1949 found its own Ratio Club (HUSBANDS; HOLLAND, 2008), under notable leadership of Ross Ashby and neurologist John Bates. In any case, Wiener continued the passage above by saying that he "found the interest in cybernetics about as great and well informed in England as in the United States, and the engineering work excellent", and completed "though of course limited by the smaller funds available."

Wiener's references to Turing, as we have seen (§1.4), included a specific recognition of the originality of Turing, who "studied the logical possibilities of the machine as an intellectual experiment" (1965 [1948], p. 13). And they also included a specific citation of Turing's results from his 1936 paper (p. 125-6) to conclude that "the logic of the machine resembles human logic, and, following Turing, we may employ it to throw light on human logic." From that Wiener proceeded to answer positively to the possibility of the machine to have even "a more eminently human characteristic," namely, "the ability to learn." In such citation, Wiener made it public that he shared Turing's non-obvious view that machines could be made to learn for themselves.

About Turing's meeting with Wiener in 1947, Andrew Hodges wrote:

> My research in the Wiener archive at MIT did not bring to light any correspondence with AMT or comment on the 1947 visit; most likely it had no great significance for either of them. (HODGES, 2012 [1983], p. 403, note 11)

As suggested by my effort with this section itself, I think that Hodges' note is far-fetched. To the least, the 1947 meeting conferred to Turing and his work a few favorable mentions in a notorious book back then. It must also have provided Turing with some inspiration or stimulus, having met someone whose views he shared to some extent. For instance, Another view of Wiener has been collected in a book on chess by a Fred Reinfeld:

> When Professor Weiner [*sic*] of the Massachusetts Institute of Technology invented a calculating machine which requires only one ten-thousandth of a second for the most complicated computations, he was quoted as saying, "I defy you to describe a capacity of the human brain which I cannot duplicate with electronic devices." (REINFELD, 2002 [1948], p. 116)

By judging from this passage attributed to him, Wiener's position seems to be *aprioristic* and laid positively towards strong mechanism. Turing's position differed from Wiener's. It was *empiricist*, as we shall see it elsewhere (§2). In any case, both Turing and Wiener had been making bold claims relative to the analogy between electronic computing machines and the human brain. They met in the spring of 1947, and their encounter rendered Wiener to make some substantial citations of Turing. Thus, I interpret, *some cross-fertilization may have taken place between them, if not in terms of ideas themselves at least in regard to inspiration or stimulus*. Yet another reason why the Wiener-Turing connection is important is that Jefferson's target in his 1949 Lister

Oration was actually Wiener (§A.4.1). But Turing shared views with Wiener, so much so that he himself replied to Jefferson, firstly in the *The Times* and then in his famous (1950) paper.

A.3.7 Turing's NPL leave of absence (Jul. 1947)

In the summer of 1947, Turing decided to ask for a leave of absence from his NPL post. He would rather spend a sabbatical year at Cambridge University. He was unhappy both with the organizational and intellectual environment at the NPL (Cf. Copeland and Proudfoot 2012, Part II) and with the slow progress on the ACE relative to rival projects in the US and in Britain. In this regard, we shall consider that in January 1947 Turing spent 20 days in a visit to the US. This let Turing identify that the Americans were ahead in several aspects in the race for the first large-scale electronic computing machine, as he reported to the NPL in his return (1947).

Essentially, Turing wanted to leave the NPL in order to be able to develop further his concept of machine intelligence. In relation to the vocabulary he used in his letter to Ashby (§A.3.4), now he was decided to explore the modeling of the higher parts of the brain, as we know from a 23 July 1947 letter by NPL director Darwin:

> As you know Dr. A. Turing [...] is the mathematician who has designed the theoretical part of our big computing engine. This has now got to the stage of ironmongery, and so for the time the chief work on it is passing into other hands. I have discussed the matter both with Womersley and with Turing, and we are agreed that it would be best that Turing should go off it for a spell.
>
> He wants to extend his work on the machine still further towards the biological side. I can best describe it by saying that hitherto the machine has been planned for work equivalent to that of the lower parts of the brain, and he [Turing] wants to see how much a machine can do for the higher ones; for example, could a machine be made that could learn by experience? This will be theoretical work, and better done away from here. The proposal then is that he should be allowed to be away for a year, which he would spend at Cambridge, where he is a fellow of King's. [...] (DARWIN, 1947)

Turing thus left the NPL and moved to Cambridge in the autumn of 1947, as determined by Copeland (2004, p. 400). He would never return. A small "pilot model" of NPL's ACE would in effect become operational only too late, in May 1950 (*Ibid.*, 367).

A.3.8 The Manchester "Baby" machine (Jun. 1948)

While in his sabbatical year at Cambridge, Turing accepted an offer from Max Newman to join the University of Manchester as a Reader in the Mathematics Department and as a "Deputy Director" of the Computing Laboratory. Newman, then the Fielden Chair of Pure Mathematics at the university, established the computer initiative for the development of the first general-purpose electronic computing machine (COPELAND, 2006; ANDERSON, 2013, p. 107; p. 31). This machine attracted Turing to the University of Manchester in the first place. It was in fact the first realization of a universal Turing machine, and the only actual general-purpose machine which

Turing would be able to experiment with. *It was also the living object that drew the attention of the press in Manchester and of Jefferson in particular, feeding his controversy with Turing*.

The project had arisen from an application that he submitted to the Royal Society in February 1946 for funding the general-purpose computer project. Crispin Rope reports (2010) that the Royal Society referred the request to Douglas Hartree (Plummer Professor of Mathematical Physics and member of the Cavendish Laboratory at the University of Cambridge) and Darwin (NPL director) to advise them. Darwin would have opposed it by arguing that the ACE at NPL would be sufficient to serve the needs of Britain, but Hartree would have recommended the grant and won the discussion against Darwin. So Hartree seems to have been a key figure for the Manchester computer project to actually exist. To actually build the machine in ironmongery, Newman recruited Professor F. C. (Freddie) Williams and his assistant Tom Kilburn at the Electrical Engineering Department. And, in connection with Hartree's recommendation to fund the project, we may note as reported by Crispin Rope that Williams had worked with Hartree before the war when they jointly constructed the automatic curve tracer for Hartree's differential analyser. According to a 1976 interview of Williams, as Copeland relates (2004, p. 209), Newman would have introduced them to Turing's 1936 idea of a universal machine, which could be implemented as a stored-program computer. Newman then explained to them what facilities were necessary in a computer. That Turing and his ideas were key to Newman's project, we know from his 8 February 1946 letter to John von Neumann:

> I am [...] hoping to embark on a computing machine section here, having got very interested in electronic devices of this kind during the last two or three years [...] I am of course in close touch with Turing. (NEWMAN, 1946)

According to Copeland (2006, p. 109-10), after resigning his NPL position Turing was appointed to his position at the University of Manchester in May 1948. This was still before 21 June 1948, when the Manchester machine became first operational and was run. At this point, the engineers were already getting credit for the advent of implementing the first universal Turing machine. On 3 August 1948, their note would appear in *Nature*:

> A small electronic digital computing machine has been operating successfully for some weeks in the Royal Society Computing Machine Laboratory, which is at present housed in the Electrical Engineering Department of the University of Manchester. The machine is purely experimental, and is on too small a scale to be of mathematical value. It was built primarily to test the soundness of the storage principle employed and to permit experience to be gained with this type of machine before embarking on the design of a full-size machine. However, apart from its small size, the machine is, in principle, 'universal' in the sense that it can be used to solve any problem that can be reduced to a programme of elementary instructions; the programme can be changed without any mechanical or electro-mechanical circuit changes. (WILLIAMS; KILBURN, 1948)

In fact, compared with the American projects to build large-scale machines, the Manchester machine was meant to be small-scale and was thus called the "Baby." Williams and Kilburn

were in fact doing cutting-edge work on storage (or "memory") technology, which became known as "Williams' tube" and was soon acquired by IBM. It turns out however that, mixed with that, was the credit for having built altogether the first universal machine, that is, the first computer to use electronic memory rather than punchcards for programming, heralding the software revolution. And in regard to that it seems that they did not share credit at the time with the mathematicians. Overall, their work started from and was guided by Newman and Turing. The mathematicians brought the foundational ideas from Turing's 1936 paper (pre-war period) and their experience with code-breaking machines (from wartime), as recent historiographies have shown (ANDERSON, 2007; COPELAND, 2011). For instance, it has been determined only later by interviews that Kilburn was actually in the audience of Turing's NPL 1947 lecture (Cf. COPELAND, 2004, p. 373). In fact, only much later Williams would acknowledge:

> Now let's be clear [...] that neither Tom Kilburn nor I knew the first thing about computers when we arrived in Manchester University. [...] Newman explained the whole business of how a computer works to us. Tom Kilburn and I knew nothing about computers. [...] Professor Newman and Mr A. M. Turing [...] knew a lot about computers [...]. They took us by the hand and explained how numbers could live in houses with addresses [...]. (COPELAND, 2006, p. 116)

Newman's take on the issue whether the "Baby" was to be credited as an invention of the engineers is commented by his son William Newman this way:

> Max had hoped that mathematicians would play a major role in computing. At Manchester, however, it was the design and construction of the computer's general-purpose stored-program hardware that took priority. Meanwhile Max gradually withdrew from the computing activity. He would explain this later by saying that 'the engineers took over', but it seems likely that his decision was influenced by his opposition to using the Manchester computer in the development of nuclear weapons. (NEWMAN, 2006, p. 187)

Newman's opposition to the development of nuclear weapons may have been influenced by their meetings with Bertrand Russell (*Ibid.*, p. 186-7), who, as known, notably manifested himself as a pacifist and against nuclear weapons in postwar Britain.

Now, what about Turing? In his (1955) memoir of Turing, Newman related:

> In 1948 he was appointed to a Readership in the University of Manchester, where work was beginning on the construction of a computing machine by F. C. Williams and T. Kilburn. The expectation was that Turing would lead the mathematical side of the work, and for a few years he continued to work, first on the design of the sub-routines out of which the larger programmes for such a machine are built, and then, as this kind of work became standardized, on more general problems of numerical analysis. (NEWMAN, 1955, p. 254)

In particular, according to Copeland (2004, p. 401), Turing designed the input mechanism and programming system for an expanded version of what ended up being Kilburn and William's

"Baby." He also wrote the *The Programmers' Handbook for the Manchester Electronic Computer* (1951). In regard to making the machine to play chess, Turing would have taught a colleague, Dietrich Prinz, how to to program the expanded version of Manchester "Baby" machine, so that it could be made to play chess (2017, p. 339).

A.3.9 Turing's NPL *Intelligent machinery* report (Summer 1948)

In the summer of 1948, between July and August woking already from Manchester, Turing submitted to the NPL a report, *Intelligent machinery* (2004 [1948]), which was expected as a result of his half-payed sabbatical year at the University of Cambridge. This text has been kept secret in the classified files of the NPL until 1968, as reported by Copeland (2004, note 53, p. 409).

About its reception at the NPL, director Charles Darwin wrote: "Dr. Turing had been on leave of absence at Cambridge University [and] has now produced a report which, although not suitable for publication, demonstrated that during his stay there he had been engaged in rather fundamental studies" (1948, p. 4). Copeland reports to have been told by Robin Gandy in a 1995 interview that Darwin would have also said that it was like a "schoolboy's essay." But as Copeland pointed out, "[i]n reality this far-sighted paper was the first manifesto of AI [artificial intelligence]" (2004, p. 401). Towards the end of his 1948 NPL report, Turing gave this summary:

> The possible ways in which machinery might be made to show intelligent behaviour are discussed. The analogy with the human brain is used as a guiding principle. It is pointed out that the potentialities of the human intelligence can only be realised if suitable education is provided. The investigation mainly centres round an analogous teaching process applied to machines. The idea of an unorganised machine is defined, and it is suggested that the infant human cortex is of this nature. Simple examples of such machines are given, and their education by means of rewards and punishments is discussed. In one case the education process is carried through until the organisation is similar to that of an ACE. (TURING, 2004 [1948], p. 431-2)

Turing's 1948 NPL report was indeed a study on machine learning based on "[t]he analogy with the human brain" in general, and with the (infant) "human cortex" in particular. This 1948 study follows through with Turing's plans as of 1946 as he had written to Ross Ashby (§A.3.4). Turing proposed a computation model that he referred to as "unorganized machines." Copeland and Proudfoot (1996) considered that it anticipated the connectionist computation model that later became known as neural networks. Turing also briefly described a sort of dual approach to machine intelligence based on his notions of "discipline and initiative." Discipline would be emphasized when a universal machine reproduces special-purpose machines strictly. Initiative would be in turn when the general-purpose machine implements, say, a connectionist model that can learn patterns of behavior directly from experience in addition to being instructed explicitly.

But in fact, Turing's 1948 NPL report goes beyond that. As we have seen (§1.6), it was in this 1948 text that Turing first articulated a list of objections to the possibility of machine thinking. Here Turing settled his definitive position (reproduced later in 1950 and 1951) on the objection to machine thinking based on Gödel's argument. He wrote (2004 [1948], p. 411): "[t]he argument from Gödel's and other theorems (objection (d)) rests essentially on the condition that the machine must not make mistakes. But this is not a requirement for intelligence."

We also have seen before (§1.2) that in this 1948 text Turing referred to "our task of building a 'thinking machine"' and to "electronic brain" (2004 [1948], p. 420). Now, I shall emphasize three core novelties of Turing's 1948 report from the point of view of this chronology.

The first key novelty is relative to intellectual fields in which one could explore and test machine intelligence. Turing had already talked about exploring the possibility of machine intelligence in the context of a game of chess before in 1945 (§A.3.2) and in 1947 (§A.3.5). But *in his 1948 NPL report Turing first presented a clear rationale about several intellectual fields and their pros and cons to suit for testing machine intelligence*. He considered five fields (2004 [1948], p. 420), including the field of "various games e.g. chess, noughts and crosses, bridge, poker;" but also "[t]he learning of languages," "[t]ranslation of languages," "[c]ryptography," "[m]athematics." We shall see his discussion of these fields elsewhere (§3.4). For now, it suffices to say that Turing had the learning of languages as "the most impressive" one, "since it is the most human of these activities." However, as it seemed to Turing as of the summer of 1948, it "depend[s] rather too much on sense organs and locomotion to be feasible." So, in connection with what comes next in the third novelty, the game of chess was preferred.

A second core innovation is concerned with Turing's concept of thinking or intelligence (he did not distinguish these words into different concepts). It turns out that Turing elaborated a new, non-obvious way to see intelligence. In connection with objection "(a)" (p. 410), that would be named in 1950 the "heads in the sand objection," Turing wrote: "the idea of 'intelligence' is itself emotional rather than mathematical" (p. 411), and in fact entitled §13 "[i]ntelligence as an emotional concept," where he characterized intelligence as a quality that is attributed according to both "our own state of mind and training" as subjective observers and "by the properties of the object under consideration." Again I shall remind the reader that my focus here is more on keeping track of Turing's ideas over time and less on their analysis. I shall study Turing's notion of intelligence as emotional in detail elsewhere (§2.3).

The third key development that Turing brought forward in his 1948 NPL report for the sake of this chronology is the very notion of an imitation game, then yet unnamed. In that summer of 1948, still unable to use the Manchester "Baby" for his experiments, Turing created in collaboration with his friend, statistician David Champernowne, the notion of a "paper machine" (2004 [1948], p. 416). The paper-machine scheme was named after them "Turochamp" (Cf. COPELAND; PRINZ, 2017, p. 331). Essentially, the paper machine was a scheme for a human being to simulate a machine *in playing chess*. In fact, the game of chess was at that time

Turing's choice of field for experiments on machine intelligence. *As of the summer of 1948, chess was Turing's preferred intellectual task to test for machine intelligence, and thus his initial experiments on the imitation game were performed by making a machine to play chess*. He wrote in (2004 [1948]): "[p]laying against such a machine gives a definite feeling that one is pitting one's wits against something alive" (p. 412). With that note, I conclude this chronology of Turing's experimental years (1939-1949).

A.4 Dialogical years (1949-1952): debating machines

As we have seen, Turing spent over two years theorizing on computing machines and more ten years experimenting with their implementation during the war and afterwards. During this relatively long time, Turing had exchanges with Newman, Church and other fellow mathematicians before the war, during the war and after. Through those years, he also interacted with electronic engineers and with cyberneticians, but not with professional philosophers. Also, he only attended to two interviews in the wake of Mountbatten's address in 31 October 1946. His focus has been actually to theorize about machines (1936-1939) and to build (a universal) machine(s) (1939-1949) towards deriving implications for the extension and limits of the machines in imitating the human mindbrain. Since June 1949, Turing expanded significantly his interlocution. He got himself involved in a public controversy in the UK on whether machines can think. It was from within the polemic that Turing made his most notable communications on machine intelligence, including his famous 1950 paper that presented to the world his imitation game or test. Turing's 1950 seminal paper, I claim, can hardly be understood if not by studying this controversy.

A.4.1 The mind-machine controversy is started (Jun. 1949)

We shall now enter into what I take to be the most influential event in the chain of events that leads to Turing's (1950) test for machine intelligence. Jefferson's Lister Oration set the stage for the polemic that provoked Turing's reaction.

Jefferson's Lister Oration (9 June 1949)

Geoffrey Jefferson (1886–1961), then Professor of Neurosurgery at the University of Manchester and fellow of the Royal Society since 1947 (WALSHE, 1961), had given on 9 June 1949 in London his Lister Oration — provocatively entitled "The mind of mechanical man" — in virtue of having received the Lister Medal for 1948 by the Royal College of Surgeons of England in recognition of distinguished contributions to surgical science. Days later on 25 June 1949, Jefferson's memorial oration would appear published as an article (1949a).

In picking out such a theme for his Lister Oration, Jefferson was motivated and informed by the research and development projects to build modern computing machines (notably the

project to build the Manchester "Baby" machine hosted at his own university, cf. §A.3.8), as well as by Wiener's *Cybernetics* published in October 1948 (cf. §A.3.6), as he himself said:

> I have to rely upon and gratefully acknowledge the assistance of Professor F. C. Williams, professor of electro-technics in my own university, and the information gleaned from Dr. Wiener, of Boston, in his entertaining book on the new science that he has christened "Cybernetics" (1948). (JEFFERSON, 1949a, p. 1108)

The note on *Cybernetics* as an "entertaining book" should not look much ambiguous, as Jefferson had let right in the beginning of his article a caveat:

> No better example could be found of man's characteristic desire for knowledge beyond, and far beyond, the limits of the authentic scientific discoveries of his own day than his wish to understand in complete detail the relationship between brain and mind — the one so finite, the other so amorphous and elusive. [...] We feel perhaps that we are being pushed, gently not roughly pushed, to accept the great likeness between the actions of electronic machines and those of the nervous system. (JEFFERSON, 1949a, p. 1105)

Jefferson may have been aware of Turing already as early as of 9 June 1949 when he gave his memorial lecture, for, as we have already seen (§A.3.6), in the *Cybernetics* Wiener cited Turing multiple times and testified on Turing's leadership on the topic of intelligent machines. For instance, Wiener also cited the engineer F. C. Williams (1965 [1948], p. 122-3). And Jefferson looked up to Williams for learning about the new machines. He reported (cf. above) to have relied upon Williams, who in fact was building the Manchester computer in a (uneasy) collaboration with Turing and Newman (§A.3.8). Jack Copeland reports (cf. 2011, p. 29) he was told by Geoff Tootill (who alongside Tom Kilburn helped Williams to build the "Baby"): "Williams, Kilburn and I [...] disliked [the term 'memory'], incidentally, as encouraging the anthropomorphic concept of 'machines that think'." Not surprisingly, Jefferson seems to have thought highly of Williams:

> To be just, nothing more than analogy is claimed by most of their constructors (some, like Professor Williams, do not go so far even as that), but there is a grave danger that those not so well informed will go to great lengths of fantasy. (JEFFERSON, 1949a, p. 1108)

That Wiener and Turing's views in general were seen as outrageous by brain-expert Jefferson, one can get an idea from a book review of Wiener's *Cybernetics* given on 23 February 1949 by a John Thurston writing for *The Saturday Review* (1949). The review was entitled "Devaluing the human brain," and run: "[i]t appears impossible for anyone seriously interested in our civilization to ignore this book," and "overlook cybernetics and its tremendous, even terrifying, implications" (p. 24). Indeed, Jefferson was an Englishman seriously interested in civilization. He took note not only of Wiener's claims but also about the existence of the "Baby" at his own university.

Jefferson discoursed beautifully. There are at least three key passages of his Lister Oration (1949a) which I shall elaborate on in connection with Turing's imitation game. One

is Jefferson's exposition of René Descartes' two means to distinguish men from machines and animals (p. 1106), which is to the best of my knowledge absolutely precise and correct. Another is the formulation of his demands to accept that *machine equals brain*, namely, in short, machines should be able to write a sonnet or a concerto because of thoughts and emotions felt (p. 1110). Jefferson's demands were made famous early on the next day of his lecture after being quoted by *The Times*, as we will see next. The third passage is Jefferson's articulation of the relationship between speech and (conceptual) thinking, which was for him where "there is the sudden and mysterious leap from the highest animal to man" (p. 1109). About Jefferson's Oration (1949a) from the point of view of this chronology, it suffices to say that it was a daunting critique of the analogy between the new electronic computing machines and the human brain. Jefferson condemned the idea that machines could think, having complained that "[w]hen we hear it said that wireless valves think, we may despair of language" (p. 1110).

Although left unmentioned by Jefferson in (1949a), Turing did not escape a warning note in the editorial of the *British Medical Journal* that opened the issue in which Jefferson's article was published. As we have seen before (§1.3), the BMJ editorial admonished Turing, "one of the mathematicians in charge of the Manchester 'mechanical brain'," that he "[p]robably he did not mean [his statement about machines and sonnets] to be taken too seriously" (1949, p. 1129). Let us now see in more detail what Turing had said to *The Times*.

Turing's interview to *The Times* (10 June 1949)

The reporter from *The Times* who covered Jefferson's lecture in London must have been puzzled by Jefferson's strong observation about computers and sonnets. Perhaps having noted Jefferson's mention of the Manchester computer (the "Baby"), he set out to get in contact with Newman's Computing Laboratory in order to get a reply from the computer experts to Jefferson's comments. It turns out, as we know from a letter of Lyn Irvine (Newman's wife) collected by Newman's son William (2012, p. 40), that Newman, who was the director of the laboratory and responsible for the project (§A.3.8), was returning from a trip to Belfast (Northern Ireland) that day, and Turing, as it seems, was the one who was available to respond by a phone interview. Making use, as we have seen before (§1.3), of his sense of humor and fine irony, Turing responded sharply to Jefferson's statements. The next day Turing's was then quoted on *The Times*. As we will see next, it was thus on 11 June 1949 that the Jefferson-Turing controversy has been publicly established.

In fact, what *The London Times* reported on 11 June 1949 as being said by Turing received the headline "Calculus to Sonnet." Given its importance to the social history of machine intelligence (known today as "artificial intelligence"), I quote it below in full:

> Mr. Turing said yesterday: "This is only a foretaste of what is to come, and only the shadow of what is going to be. We have to have some experience with the machine before we really know its capabilities. It may take years before we settle down to the new possibilities, but I do not see why it should not enter

> any one of the fields normally covered by the human intellect, and eventually compete on equal terms".
>
> "I do not think you can even draw the line about sonnets, though the comparison is perhaps a little bit unfair because a sonnet written by a machine will be better appreciated by another machine".
>
> Mr. Turing added that the University was really interested in the investigation of the possibilities of machines for their own sake. Their research would be directed to finding the degree of intellectual activity of which a machine was capable, and to what extent it could think for itself. News of the experiments was disclosed by Professor Jefferson in the Lister Oration reported in *The Times* yesterday. (TIMES, 1949a)

By mediation of the reporter, Turing was responding to what on 10 June 1949 *The London Times* reported (under headline "No mind for a mechanical man" 1949b) Jefferson to have said in his Lister Oration (9 June), as also appeared days later in his text (25 June):

> A machine might solve problems in logic, since logic and mathematics are much the same thing [...]. But not until a machine can write a sonnet or a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that machine equals brain — that is, not only write it but know that it had written it. No mechanism could feel (and not merely artificially signal, an easy contrivance) pleasure at its successes, grief when its valves fuse, be warmed by flattery, be made miserable by its mistakes, be charmed by sex, be angry or miserable when it cannot get what it wants. (JEFFERSON, 1949a, p. 1110)

Thus Jefferson had posed his demands in order to accept that "machine equals brain" or to accept that a machine can think. And those words of Jefferson's were in effect key, as they influence Turing's (1950) paper over a year later to the point of having been quoted and addressed directly.

Aftermath and Lyn Irvine's letter (24 Jun. 1949)

I want to give the reader a vivid impression on the (the Newmans) reception of Turing's interview to *The Times*, and again (cf. §1.5) to address that I trust a testimony from Max Newman's wife Lyn Irvine. It is a letter that she wrote to her friend Antoinette Esher on 24 June 1949:

> Did you see the extraordinary report in the *Times* two weeks ago on the Manchester Calculating Machine with the fantastic remarks attributed to Alan Turing? And Max's letter the following week trying to clear things up? The *Times* wired Alan, who isn't on the telephone, to ring their office, and they interviewed him on the phone. He's wildly innocent about the ways of reporters and has a bad stammer when he's nervous or puzzled. It was a great shock to him when he saw the *Times* — and to Max who had been flying back from Belfast that day. We had a wretched weekend starting at midnight on the Friday night when some subeditor of a local paper rang up to get a story. By Sunday Max was getting a bit gruff, and when he said, 'What do you want?' to one newspaper, the reporter replied, 'Only to photograph your brain'. (NEWMAN, 2012, p. 40)

Actually, let us consider how original, articulate and witted *The Times*'s quotation of Turing is. It has a satirical vein that is typical of Turing, as we know from virtually all related comments in

Turing sources (§1.5). I think that it is unlikely that the reporter could have made any editing over Turing's words. Yet Turing may, of course, have manifested some half-serious regret to Newman and his wife, since the trouble his interview posed to them was inadvertent. Lyn Irvine was fond of Turing, as we know, *e.g.*, from recent recollection by her son William Newman (2012). And it seems that her perception has been weighed by the whole situation. In any case, Newman would sort it all out by writing up a sober clarifying letter on electronic computing machines to *The Times* that was published on 14 June 1949, and a note to the *British Medical Journal* that appeared in time in the same issue of the full text of Jefferson's Lister Oration.

Max Newman's note to the *British Medical Journal* (BMJ) (25 Jun. 1949)

Newman's note to the BMJ (1949) was diplomatic. It provided some backup for Turing, on the one hand, and lowered the tone and trying to emphasize the mathematical applications of the Manchester Baby as a universal machine, on the other hand. Also in that note, he provided what I have been calling Newman's distinction on the universal and the existential variants of the question on whether machines can think (cf. Introduction). He suggested that too much of the public attention was being given to the universal variant of the problem. But what was Newman's own view? Indeed, we may know from the same 24 June 1949 letter from his wife Lyn Irvine to her friend Antoinette Esher that is reproduced in Jonathan Swinton's Turing biography (2019, p. 79). Irvine wrote:

> To hear [Max] & Alan quietly chatting about the machine in our garden sometimes make me feel very queer & uncomfortable. Max hopes to teach it to see a joke. It will learn how to play chess probably in the Autumn & Max is to have the first game & a wager that he will win. That's all right [...] but when I heard Alan say of further possibilities "Wh-wh-what will happen at that stage is that we shan't understand how it does it, we'll have lost track", I did find it a most disturbing prospect. It was Professor Jefferson's Lister Oration which broke the peaceful silence & started first *The Times* & then all the tabloid papers on their stampede for stories & pictures. (SWINTON, 2019, p. 79)

Given that testimony, it cannot be the case that Newman had the sober view and Turing the fabulous one. Rather, Newman kept his view to his close circle while Turing spoke out his mind.

And about Newman, the BMJ (1949) editorial run:

> Mechanical calculating machines have existed for many years, and now, Professor M. H. A. Newman, F.R.S., informs us on another page, can be coded to play "a legally correct game of bridge, poker, or chess." To say that, because a machine can be constructed to symbolize certain processes of thought, machines will one day be made which could fall in love, write sonnets, or rule their makers is merely bad logic. (BMJ, 1949, p. 1129)

The reader may observe that, as cited by the BMJ editorial, Newman referred — just as Turing had himself done in his 1948 NPL report — to games such as "bridge, poker or chess" as a field

in which to test the new machine capabilities. As can be observed in the following sentence of the editorial, the inclusion of such games within the machine capabilities did not sound offensive. Thus after referring to "bad logic," the editorial immediately turned to "Mr. A. W. [*sic*] Turing." Indeed Turing had made a satire of Jefferson's observation about machines and sonnets (§A.4.1). But as we have seen before (§1.3), the BMJ editorial tried to push a view of his satire as not serious. In the editorial, Newman was portrayed as the informant on the state of the art of the new computing machines, while Turing was the one to receive postwar reprimand.

Follow-up events in 1949 Manchester

The Jefferson-Turing controversy is an event in the social history of the philosophy of mind and the field that we call today artificial intelligence. For a short version, I think that it can be represented by the two quotations above from *The Times* — one of Turing, the other of Jefferson. I think that the birth of that philosophical dispute can be dated to 11 June, the next day to when Turing had an opportunity to be more diplomatic (say, like Max Newman) but decided instead to push back. And the polemic went on through the year of 1949, as described next.

A.4.2 The Manchester seminars (Oct. – Dec.1949)

On 27 October 1949 a crucial event was held in the Department of Philosophy of the University of Manchester, the seminar on "the mind and the computing machine." It was co-chaired by Dorothy Emmet and Michael Polanyi (HODGES, 2012 [1983], p. 414-5). We know of the seminar mostly from minute notes that were taken, according to Wolfe Mays, by some unidentified member of the Philosophy Department. According to Jonathan Swinton, the author may have been Desmond P. Henry, then a junior member of the Philosophy Department (2019, p. 92, note 177). In any case, a copy of the notes would have been kept in possession of Dorothy Emmet and be available in her archive ("Papers of Dorothy Emmet") at Lucy Cavendish College, Cambridge (*Ibid.*). An identical copy was in possession of Wolfe Mays, who was also in the seminar then as a young lecturer of the Philosophy Department. Mays commented the notes in (2000). To my knowledge, it was only in 2005 that he made them available to Turing scholars. Jack Copeland published a facsimile in the Turing Digital Archive and a transcript in (2005 [1949]). The October 1949 seminar had two sessions. The first was led by Polanyi, and the second by Emmet. I shall now refer to the discussion that took place in the sessions, and then refer to the evidence that a second meeting was held in December 1949. (Of this second edition of the Manchester seminar, no record is yet known.) We shall start with the session driven by Polanyi.

Michael Polanyi's session (Oct. 1949)

In the first session, Polanyi read a text that he had prepared and circulated to Newman and Turing several weeks before the meeting. According to Swinton (2019, p. 91, note 176), this can be derived from a letter from Newman to Polanyi on 19 September 1949 that is in the Michael

Polanyi Papers at the Regenstein Library of the University of Chicago, Box 5, Folder 6. Polanyi scholar Paul Blum published Polanyi's text in (2010). (The original is in the same archive.)

Polanyi's session-opening text is entitled "Can the mind be represented by a machine? Notes for discussion on 27th October 1949." Looking at Polanyi's printed copy at the University of Chicago archive, Blum observed annotations that could have been made by Turing. Blum wrote (2010, p. 52) that there were a few corrections that were "certainly by Polanyi," but there were "three comments by a different hand," and speculated: "compared with manuscripts published at the Turing Digital Archive (www.turingarchive.org) they could be by Turing." I reproduce below what those three comments (possibly by Turing) are:

> The discoveries of Gödel (1930) have shown that arithmetic and advanced geometry are incomplete. [Ed. Superscript by unknown hand: *rather: number theory*.] (BLUM, 2010, p. 52, note 40)
>
> There is established thus an inexhaustible procedure for the discovery of ever more true mathematical formulae, which, by its very nature, is incapable of formalisation. [Ed. "nature... formalization": underlined and annotated by unknown hand: *no./not in the same language. But we can formalize the meta-language.*] (BLUM, 2010, p. 52-3, note 41)
>
> Our minds however are not similarly limited. [Ed. Superscript by unknown hand: *But they are. Otherwise we get into the paradoxes*.]
> (BLUM, 2010, p. 53, note 44)

In fact, these annotations are consistent with what Turing is reported to have said in the seminar itself and also with statements of his elsewhere. Let us see. About Polany's text, Newman and Turing are reported to have said:

> NEWMAN TO POLANYI: The Gödel extra-system instances are produced according to a definite rule, and so can be produced by a machine. The mind/machine problem cannot be solved logically; it must rest on a belief that a machine cannot do anything radically new, to be worked on experimentally. The interesting thing to ask is whether a machine could produce the original Gödel paper, which seems to require an original set of syntheses.
>
> TURING: emphasises the importance of the universal machine, capable of turning itself into any other machine.
>
> POLANYI: emphasises the Semantic Function, as outside the formalisable system. (TURING et al., 2005 [1949])

In fact, Polanyi thus lent to us the opportunity to see a most spontaneous reply by Newman and Turing as preserved in the seminar notes. First, it is evidence that Newman considered, just like Turing (§3.5), that "the mind/machine problem" can be decided empirically and only empirically. That is, for Newman as well, it is not merely a language issue as is sometimes suggested (cf. §3.2). More than that, Newman abstracted the problem of producing "the original Gödel paper" as the question whether a machine "can do anything radically new." *Newman therefore shifted the discussion around Gödel's argument from the mathematical objection to Lady Lovelace's objection*. Turing's reply in turn adds to his standard reply to the mathematical objection as we

know it from his 1947 and 1948 communications. In particular, he again correlated it with his take in his 1938 thesis, in connection with the notions of intuition and ingenuity, proof-checking and proof-finding machines, and most notably with the idea of an ordinal logic as analogous to a universal Turing machine (§A.2.2). *For Turing, the universal machine is capable of turning itself into any other (proof-finding) machine. In effect, Turing did not rule out the possibility for a machine to produce an argument such as Gödel's.*

I shall refer to Polanyi's *Personal knowledge* (1974 [1958]) in order to collect key secondary-source evidence for the sake of this chronology of Turing's thoughts on machine intelligence. Before further comments, let me promptly quote it:

> A. M. Turing has shown [Polanyi's note: in a communication to a Symposium held on "Mind and Machine" at Manchester University in October, 1949. This is foreshadowed in 'Systems of Logic based on Ordinals', *Proc. London Maths. Soc.*, Series 2, 45, 1938-9, pp. 161-228.] that it is possible to devise a machine which will both construct and assert as new axioms an indefinite sequence of Gödelian sentences. Any heuristic process of a routine character — for which in the deductive sciences the Gödelian process is an example — could likewise be carried out automatically. A routine game of chess can be played automatically by a machine, and indeed, all arts can be performed automatically to the extent to which the rules of the art can be specified. While such a specification may include random elements, like choices made by spinning a coin, no unspecifiable skill or connoisseurship can be fed into a machine.
> (POLANYI, 1974 [1958], p. 261).

So Polanyi thereby related that, like him, Turing also made a communication in the first session of the October 1949 seminar. Also, Polanyi suggested that what Newman is reported to have said in the previous quotation had been already said earlier in the same meeting by Turing. Third, it provides evidence that Turing's reference to "the universal machine, capable of turning itself into any other machine" (cf. above) is connected with Turing's 1938 paper, indeed. (It was connected from the point of view of Polanyi, and I doubt that he could have made this connection himself if not by recalling Turing's own seminar presentation.) Fourth, it shows that as of October 1949 Turing was still referring to the game of chess as intellectual task to showcase the possibility of machine intelligence.

Dorothy Emmet's session (Oct. 1949)

In the second session, Emmet outlined three "questions to be considered" for discussion, namely, (1) "Machine brain analogy," (2) "Physiological aspects," and (3) "Are there any limitations to the kind of operations which a machine can do?" By looking at the recorded notes (TURING et al., 2005 [1949]), I observe that Turing engaged with the discussion of questions 1 and 3. They were focus at the beginning and at the end of the discussion. Let us look at Turing's interventions.

At the beginning, Emmet would have explicitly asked about machines and purpose. Turing would have given this non-obvious reply:

> Questions asked: Is it possible to give a purpose to a machine? Can you 'put a purpose into a machine'?
>
> TURING: This kind of thing can be done by 'trial and error' methods: purpose is 'use of previous combinations plus trial and error'.
> (TURING et al., 2005 [1949])

About Turing's comment, Mays wrote in (2000, p. 62): "(Turing was obviously thinking here of feedback mechanisms, sometimes called goal-seeking devices, which by trial and error gradually approach the target or goal. [...])." Mays mixes up Turing's views with Wiener's cybernetics. I do not think that it is all wrong, as they are not totally unrelated. But Mays' appeal to obviousness is handwavy. Turing's view is specific and deserves to be considered in its subtleties, in connection to Turing's 1948 notion of "initiative" and his related 1950 notion of learning machines.

Following on in Emmet's session the discussion was then suddenly shifted to brain physiology after interventions by Maurice S. Bartlett (a statistician) and J. Z. Young (the famous physiologist). Next to a relatively long discussion that moved on to scientific explanation and the role of models in science, Hewell (a researcher at Manchester-based pharmaceutical Imperial Chemical Industries, or ICI) mentioned the word "choice," which brought Turing back to the discussion with an observation about "random operation" (as a way to, say, simulate choice). Soon then Turing would have presented a distinction, to which Polanyi replied:

> TURING: declares he will try to get back to the point: he was thinking of the kind of machine which takes problems as objectives, and the rules by which it deals with the problems are different from the objective. Cf. Polanyi's distinction between mechanically following rules about which you know nothing, and rules about which you know.
>
> POLANYI: tries to identify rules of the logical system with the rules which determine our own behaviour, and these are quite different things.
> (TURING et al., 2005 [1949])

Here the chances were high that we would get back to the topic of the first session. But Emmet avoided that with a key intervention that lent us to have Turing developing further his first-session reply to Newman and Polanyi:

> EMMET: the vital difference seems to be that a machine is not conscious.
>
> TURING: a machine may act according to two different sets of rules, e.g. if I do an addition sum on the blackboard in two different ways:
>
> 1. by a conscious working towards the solution
> 2. by a routine, habitual method
>
> then the operation involves in the first place the particular method by which I perform the addition – this is conscious: and in the second place the neural mechanism is in operation unconsciously all the while. These are two different things, and should be kept separate. (TURING et al., 2005 [1949])

So Turing replied that the act of choosing a method to accomplish a task can be seen as the exercise of consciousness. Also implied in that reply, I interpret, Turing thought that machines

can afford self-referential operations. He would develop that notion further in his 1950 paper when discussing objection 5 ("arguments from various disabilities").

At that point Polanyi made once more a reference to his notion of "semantic function." I quote it below together with Turing's reply:

> POLANYI: interprets this as suggestion that the semantic function can ultimately be specified; whereas in point of fact a machine is fully specifiable, while a mind is not.
>
> TURING: replies that the mind is only said to be unspecifiable because it has *not yet been* specified; but it is a fact that it would be impossible to find the programme inserted into quite a simple machine – and we are in the same position as regards the brain. The conclusion that the mind is unspecifiable does not follow.
>
> POLANYI: says that this should mean that you cannot decide logical problems by empirical methods. [...] (TURING et al., 2005 [1949], no emphasis added)

In effect, Polanyi would only describe what the semantic function underlying the human mind is in his 1958 book. What is most interesting here is to consider Turing's spontaneous reply to it. This discussion must have been the source of Turing's 1950 formulation of (objection 7) the argument from informality of behavior, which we have seen before (§1.6). As for Polanyi's last statement quoted above, I shall say that it was a bright philosophical observation. It can be seen, in effect, as intuition that the discussion was lacking a crucial element. That missing element was an *epistemological standard for thinking*. Turing could have brought it forward by referring to his own communication in the first session of the seminar and to the game of chess. He did not. Perhaps chess would sound too low profile to serve as reference in Emmet's session discussion.

Evidence of a December 1949 edition of the seminar

We also know of another edition of that seminar to have taken place again in the Philosophy Department in December 1949. Let us recall Jefferson's reminiscence of Turing that I quoted before (§1.2). He remembered of Turing having come to his house to talk to Professor J.Z. Young and him after a meeting in the Philosophy Department arranged by Dorothy Emmet. The key information that Jefferson gave was that Turing went there and left on his bicycle "through the *winter* rain" (emphasis added). So that meeting cannot have been the October 1949 one (whose minute notes survived). It must have been another meeting in late December or early in 1950. Now, Jefferson's reference matches Swinton's finding (1949) of a Christmas Eve postcard sent to Warren McCulloch then in Chicago by Jules Bogue, who according to Swinton was an American running the drugs division of Manchester-based company Imperial Chemical Industries in Wilmslow and that happened to have been neighbor of Max Newman, and was in the meeting. Thus run the postcard:

> I wish you [McCulloch] had been with us a few days ago we had an amusing evening discussion with Thuring [*sic*], Wiliams, Max Newman, Polanyi,

> Jefferson, JZ Young & myself. An electronic analyser and a digital computer (universal type) might have sorted the arguments out a bit. (BOGUE, 1949)

It is worth noting the name of F. C. Williams as present in the December edition of the Manchester seminar. Some confusion in the discussion is noted.

Towards Turing's *Mind* 1950 paper

Altogether, leading to Turing's *Mind* 1950 paper, there is a key point that I am concerned with in this chronology of Turing's thoughts on machine intelligence, namely, the intellectual field(s) that Turing has considered and chosen for his imitation game. We have seen related from Polanyi that, in the October 1949 seminar, Turing would have made a communication in which he referred to chess convincingly enough to make Polanyi write: "A routine game of chess can be played automatically by a machine, and indeed, all arts can be performed automatically to the extent to which the rules of the art can be specified" (1974 [1958], p. 261). I take that as showing that as of the opening of the October 1949 seminar, chess was still Turing's preferred choice of intellectual task to illustrate and test for machine intelligence. So, since Turing's very first thoughts on machine intelligence during the war through Turing's 1945, 1947 and 1948 NPL reports and lectures *up to the opening of the 1949 Manchester seminars, Turing still had the game of chess as his preferred choice of intellectual task to discuss and test for machine intelligence*. But this would change.

A.4.3 Turing's *Mind* paper (*c.* early 1950)

Turing's (1950) paper was published in October 1950. It must have been early, at some point from the winter (after the December 1949 seminar) to the spring of 1950, that Turing wrote and submitted his text to *Mind*. We know that also from a comment by Mays (2001, p. 4): "In the summer of 1950 Gilbert Ryle sent me the galley proofs of Aean [*sic*] Turing's article 'Computing Machines and Intelligence' [*sic*] which was to appear in the October number of *Mind* 1950." If Simon Lavington's timeline is correct in its 8 February 1950 entry (2012, p. 100, no actual source is given) such that on that day Turing wrote that he was already working on his 'mathematical theory of embryology' (morphogenesis), then he most likely wrote his 1950 paper in *c.* January 1950. Indeed, we also know from Robin Gandy, as pointed out by Copeland (2004, p. 433), that Turing would have written it "unlike his mathematical papers quickly and with enjoyment" (1996, p. 125).

And yet Turing's 1950 paper is the high moment of development of his views how to test for machine intelligence. I will offer a dedicated analysis of it later (§3). For here, a general outlook of Turing's paper will be helpful with focus on what was new in it with respect to machine intelligence as a concept that Turing had actually been developing since 1936. Let us explore the main novelties on machine intelligence.

In the very opening of his text, *Turing made a point about the dispensability of defining words "machine" and "thinking" according to the common sense back then*. He addressed the issue of the paradoxical aspect of the combining words "machine" and "thinking" or "intelligence" directly. We shall see it in detail (§3.3). As we have seen in the beginning of this chapter, it did not touch, for example, Wolfe Mays, who must have found that no Gallup survey was even needed since he could just quote from the dictionary. The problem of committing to a definition of thinking will be posed to Turing directly by Jefferson in 1952, as we shall see later (§A.4.6).

Turing also dedicated a part of his text to teach his new science of computing. This does not add to what we have seen before (§A.2) and yet its very existence is relevant (from a structural point of view) for this chronology. In the transition to his discussion of objections, having reviewed the state of the art of digital computers, Turing wrote (1950, p. 442): "[w]e may now consider the ground to have been cleared and we are ready to proceed to the debate on our question, 'Can machines think?' and the variant of it quoted at the end of the last section." I take that as evidence that Turing viewed his discussion (his sections §§6, 7) as dependent on his teachings about machines. That is, *for Turing, both his negative and positive argumentation are grounded on his science of computing and the existence of digital computers as universal computing machines.* This also adds strength to the view that Turing, as a philosopher, tried to keep his thoughts tied up to his science.

In the opening of his section §6, Turing outlined his beliefs relative to the question whether machines can think. He offered *two empirically testable predictions* for the future. His science-and-technology prediction stated that in about fifty-years time it would be possible to program a machine (it would have enough storage capacity) to play the imitation game well. Turing's social-and-cultural prediction stated that although the original question "[c]an machines think?" was too meaningless to deserve discussion, the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. The other novelties of Turing's section §6 are concerned with a few updates to the five 1948 objections and with the inclusion of four new objections (numbers 4, 7, 8 and 9; cf. Figure 1). Three of them have dialogical connections with Polanyi (the argument from informality of behavior) and Jefferson (the argument from consciousness and the one from continuity in the nervous system). I will discuss them in detail later (§3.7). Overall, Turing's 1950 discussion was able to streamline much of the previous debate on topics such as consciousness and continuity in the nervous system. In 1950 Turing proposed conversational question-answering as an epistemological standard for thinking, actually borrowed from Jefferson's Lister Oration.

Novelties of Turing's section §7 in relation to the imitation game will also be analyzed later (§3.3). Also, Turing gave as of 1950 a dominating image, the "skin of an onion" analogy, to explain the possibilities he considered on the extension and limits of the machine-mindbrain analogy. Some of the operations of the mind or the brain, Turing wrote, can be explained in purely mechanical terms. As to whether the same holds for the whole of it is yet an open problem to be

investigated scientifically. One may proceed gradually, as if stripping off the multiple layers of skin of an onion, if at the end there is nothing else, then the whole mind is mechanical. Essentially, with that move, Turing clearly positioned that, on the one hand, he was not considering any *a priori* limits on the machines' intellectual capabilities relative to ours as human beings; and on the other hand, he was not committed beforehand to viewing their intellectual power as unlimited either. For Turing, the question deserved more studies indeed. I will study Turing's position about the mind and intelligence later (§2). Finally, there were two core developments in Turing's 1950 section §7 in connection with the concept of machine intelligence. First, he claimed that the problem is mainly one of programming. Advances in engineering would have to be made too, but for Turing it was unlikely that these would not be adequate for the requirements. Second, in comparison to his 1948 NPL report, Turing advanced his views on the problem of how to educate a "child machine." He outlined a more detailed research strategy for it. He suggested a balanced approach between two sources of learning, one based on binary outcomes of situations the machine experiences, and the other based on an imperative, symbolic yet uncertain inference (not meant to "satisfy the most exacting logicians"). Turing's proposed methods in Step 4 seem to better materialize what he had called in 1948 "discipline and initiative."

Overall, as I have developed before (§1.5), Turing's *Mind* paper is the occasion when he clearly assumed his role as prophet of the machines. Let us now proceed to his follow-up discussion on machine intelligence from 1951 to 1952.

A.4.4 Turing's BBC radio lecture "Intelligent machinery" (*c.* 1951)

It was at some point in the year of 1951, according to Copeland (2004, p. 465), that Turing would have given the lecture entitled "Intelligent machinery, a heretical theory" on a radio discussion program called *The '51 Society*. The program would have been named after the year it first went to air, and was produced by the BBC Home Service at their Manchester studio (*Ibid.*, p. 465). A copy of the typescript is kept in the Turing Papers at King's College, Cambridge. A transcript has been first published by Copeland (1999). I quoted from Turing's *c.* 1951 radio lecture before in my assessment of his (Promethean) ambition (§1.5). There, we have seen, Turing said that intelligent machines "are a genuine possibility," and that "it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers" (2004 [*c.* 1951], p. 475). I have claimed that, although there was irony in Turing's words, he meant that the possibility of (super)intelligent machines is germane indeed. So this is a first key topic that updates this chronology. Let us now consider other novel statements made by Turing in his *c.* 1951 lecture.

It is a short communication, whose transcript has only seven paragraphs. Turing opened it by promptly declaring his intention to challenge conventional wisdom (paragraph 1):

> 'You cannot make a machine to think for you.' This is a commonplace that is usually accepted without question. It will be the purpose of this paper to

> question it. (TURING, 2004 [*c.* 1951], p. 472)

The next six paragraphs can be organized in three logical steps: first he provided a new variant of his discussion of *the mathematical objection* (paragraphs 2 and 3); then proceeded to bring forward a new variant of his discussion of *learning machines* (paragraphs 4, 5 and 6); and finished it with his comment on the *possibility of (super)intelligent machines* (paragraph 7).

So, first Turing sheds new light on how he sees machines in comparison to human mathematicians in terms of intellectual capabilities. He discussed Gödel's argument and its variations by which, as he has taken, "one can show that however the machine is constructed there are bound to be cases where the machine fails to give an answer, but a mathematician would be able to" (p. 472). Turing, nonetheless, pushed back the burden of proof, having said: "but this cannot be regarded as very different from the reaction of the mathematicians, who have for instance worked for hundreds of years on the question as to whether Fermat's last theorem is true or not." He then puts it most clearly:

> My contention is that machines can be constructed which will simulate the behaviour of the human mind very closely. They will make mistakes at times, and at times they may make new and very interesting statements, and on the whole the output of them will be worth attention to the same sort of extent as the output of a human mind. (TURING, 2004 [*c.* 1951], p. 472)

Thus he shared what his beliefs were. So he added to his 1950 "skin of an onion" analogy another statement as of *c.* 1951 to insist that he was not considering any *a priori* limits for the machines' intellectual capabilities relative to ours as human beings.

In his reiterated discussion of learning machines, like in 1950, Turing emphasized that what he had in mind was to make the machine to "learn by experience." About the machine education process, and how it could be programmed, Turing made a discussion not very different from the related on he made in (1950). He discussed the possibility of building for the machine "indexes of experiences" to catalog a range of possible situations, the use of pleasure-pain keys by the schoolmaster, and the inclusion of random element to simulate selection, choice or initiative, and not let "the behavior of the machine [to be] completely determined by [its] experiences."

Also in view of the reception of Turing's proposals, there are two aspects that I want to point out. First, let us consider Turing's comment about the possibility of cheating, or "having a man inside the machine:"

> This might be called 'education'. But here we have to be careful. It would be quite easy to arrange the experiences in such a way that they automatically caused the structure of the machine to build up into a previously intended form, and this would obviously be a gross form of cheating, almost on a par with having a man inside the machine. (TURING, 2004 [*c.* 1951], p. 473)

I take from this passage that, *for Turing, the capability of learning for itself is a hard requirement for the machine to be considered intelligent, and cheating is not allowed*. (This leaves open the non-obvious question of what could or should be considered cheating. I will come back to this topic in §2.4). Also relevant for this chronology is this next passage where Turing wanted to make a caveat:

> To make further suggestions along these lines would perhaps be unfruitful at this stage, as they are likely to consist of nothing more than an analysis of actual methods of education applied to human children.
> (TURING, 2004 [*c.* 1951], p. 475)

I interpret that *Turing attributed to his specific suggestions about how to educate the machine a status of preliminary research strategies*. They were made as nothing but a best effort of his based on what he knew about education back then.

A.4.5 Turing's BBC lecture "Can digital computers think?" (May 1951)

In 15 May 1951, according to Copeland (2004, p. 476), Turing's radio lecture entitled "Can digital computers think?" was first broadcast on BBC radio. It was the second in a series with the general title 'Automatic Calculating Machines.' Other speakers in the series included Max Newman, Douglas Hartree, M. V. Wilkes, and F. C. Williams, all computer pioneers as well. More context about the BBC series by paying attention to the other participants and their contributions to the series is given by Allan Jones (2004). A copy of the typescript is kept in the Turing Papers at King's College, Cambridge, and it was first published by Copeland (1999). I quoted from Turing's May 1951 BBC lecture before in my assessment of his (Promethean) ambition (§1.5). There, as we have seen, Turing said that "this new danger" of having machines superseding us in intelligence power "is much closer," and continued: "it is remote but not astronomically remote, and is certainly something which can give us anxiety" (2004 [1951], p. 486). He then declared, as I interpreted, his positive humanistic hopes in the study of the extension and limits of machine intelligence, and his dislike of post-humanistic projects. From the point of view of this chronology this only adds strength to the view that, as we have just seen, Turing thought that (super)intelligent machines are a true possibility.

Let us now consider core novelties brought forth by Turing in his May 1951 radio lecture. In fact, like the *c.* 1951 one, this was also a short communication. Its transcript is structured in seventeen paragraphs. Again, Turing opened it strikingly (paragraph 1):

> Digital computers have often been described as mechanical brains. Most scientists probably regard this description as a mere newspaper stunt, but some do not. One mathematician has expressed the opposite point of view to me rather forcefully in the words 'It is commonly said that these machines are not brains, but you and I know that they are.' (TURING, 2004 [1951], p. 482)

He concluded this first paragraph by saying that his goal is "to explain the ideas behind the various possible points of view, though not altogether impartially." He would focus on his own view, "that it is not altogether unreasonable to describe digital computers as brains."

Strongly coupled to this first observation, the next three paragraphs presented, respectively, (2) the view of the man in the street, (3) the view of scientists, represented by Lady Lovelace and Douglas Hartree, and (4) his own. Turing considered that the man in the street is bound to be persuaded (*cp.* with his 1950 social-and-cultural prediction) of the intelligence of the new machines, as they are capable of "intellectual feats of which he would be quite incapable." (p. 482). The scientists, on the other hand, tend to be contemptuous of that as a superstitious attitude. They seemed to be eluded, Turing considered, by how digital computers were actually used at their time, and by how "they will probably mainly be used for many years to come." He then contended, exposing his own point of view, that he agreed with Lovelace's dictum that computers "can do whatever *we know how to order it* to perform" (no emphasis added, p. 482). But its validity depends, Turing said, "on considering how digital computers *are* used rather than how they *could* be used" (emphasis is Turing's). He added: "[i]n fact I believe that they could be used in such a manner that they could appropriately be described as brains" (p. 482). The next paragraphs (5-17) of Turing's radio lecture can be organized in six other steps, as described next.

In Turing's Step 2 (paragraphs 5-8), core assumptions and properties of digital computers are discussed: their universality, their suitability to be programmed to imitate any other machine, the meaning of the "mechanical" analogy with brains, and their storage capacity.

Moving to Step 3 (paragraphs 9-11), Turing took a moment to position the relative roles of what we would call today hardware and software in making a digital computer to imitate a brain. And here is when it comes what is in my view the central point of Turing's May 1951 radio lecture as an argument in support of describing digital computers as brains:

> It should be noticed that there is no need for there to be any increase in the complexity of the computers used. If we try to imitate ever more complicated machines or brains we must use larger and larger computers to do it. We do not need to use successively more complicated ones. This may appear paradoxical, but the explanation is not difficult. The imitation of a machine by a computer requires not only that we should have made the computer, but that we should have programmed it appropriately. The more complicated the machine to be imitated the more complicated must the programme be.
> (TURING, 2004 [1951], p. 483)

That is, *for Turing*, let me insist, *as long as enough storage capacity is provided, the bottleneck to make a digital computer to imitate a brain is in the software, not in the hardware*. Although not really new in this chronology, as Turing had already stated it in 1950, in this 1951 radio lecture he significantly developed it. In fact, this is a key point in connection with received views, specially with proponents of the mathematical objection such as Michael Polanyi (who after Newman and Turing's replies in the October 1949 seminar shifted his focus to the argument

from consciousness), John Lucas and Roger Penrose. In Step 3 still, Turing made this other key remark about what would be, for him, the "wisest ground on which to criticize" his views:

> In view of this it seems that the wisest ground on which to criticise the description of digital computers as 'mechanical brains' or 'electronic brains' is that, although they might be programmed to behave like brains, we do not at present know how this should be done. With this outlook I am in full agreement. It leaves open the question as to whether we will or will not eventually succeed in finding such a programme. I, personally, am inclined to believe that such a programme will be found. (TURING, 2004 [1951], p. 484)

Turing then restated his 1950 belief that "it is probable" that "at the end of the century" it would be possible to program a machine to perform well in what he described in 1950 as the imitation game. He also restated the format of the imitation game as a *viva-voce* examination.

Now in Step 4 (paragraphs 12-13), Turing wanted to acknowledge "some difficulties," namely, "[t]o behave like a brain," he said, "seems to involve free will, but the behavior of a digital computer, when it has been programmed, is completely determined." Turing so acknowledged to find himself within "an age-old controversy:" free will and determinism. He thus sorted it out:

> There are two ways out. It may be that the feeling of free will which we all have is an illusion. Or it may be that we really have got free will, but yet there is no way of telling from our behaviour that this is so. In the latter case, however well a machine imitates a man's behaviour it is to be regarded as a mere sham. I do not know how we can ever decide between these alternatives but whichever is the correct one it is certain that a machine which is to imitate a brain must appear to behave as if it had free will [...]. (TURING, 2004 [1951], p. 484)

For Turing free will is either an illusion or unobservable. In the latter case it would be forever reserved by dogma to human beings only, so there would be nothing one could do about it — "however well a machine imitates a man's behaviour it is to be regarded as a mere sham." In any case, Turing's conclusion is that a machine supposed to imitate a brain "must appear to behave as if it had free will." He discussed the possibility of introducing a random element into the machine, but then dismissed it as a hard requirement. Included it or not, Turing said, the machine's behavior may appear random anyway to someone who is not familiar with its program — just like a roulette wheel or a supply of radium may somehow be predictable, and yet we may not know how to make the prediction. For "a machine to imitate a brain, or as we might say more briefly, if less accurately, to think," Turing added, "[w]e must not always expect to know what the computer is going to do." Turing further clarified it by analogy: "we should be pleased when the machine surprises us, in rather the same way as one is pleased when a pupil does something which he had not been explicitly taught to do" (p. 485). Turing suggested overall, I interpret, that *whatever its internal mechanisms are, including random elements or not, for a machine to be regarded as thinking it must be capable to surprise us*.

Turing was then ready to revisit in his Step 5 (paragraph 14) the question of Lady Lovelace's dictum, "[t]he machine can do whatever *we know how to order it* to perform" (p. 485,

no emphasis added). He observed that one (as he mentioned earlier, scientists such as Hartree) may be tempted to add "only" to it, holding that the machine can *only* do what we know how to order it to perform. Turing argued that there is no need to suppose that when we give orders to the machine we know exactly what we are doing, or what the consequences of these orders are going to be. This is not so, he said, "any more than one needs to understand the mechanism of germination when one puts a seed in the ground." Turing completed: "[t]he plant comes up whether one understands or not."

In Step 5, just like he did in the *c.* 1951 lecture, Turing once more emphasized that he is not actually committed with a specific education process for the machine, as long as it "bear[s] a close relation of that of teaching" (p. 485). Finally, in his last step, as we have already seen (§1.5), Turing outlined his views on the possibility of (super)intelligent.

A.4.6 The BBC roundtable "Can…machines…think?" (Jan. 1952)

On 10 January 1952 there was a roundtable discussion at BBC on the question "[c]an automatic calculating machines be said to think?" The discussion was aired first on 14 January, and again on 23 January. Besides Turing, Newman and Jefferson, the BBC invited Richard Braithwaite (1900-1990). He was then Sigwick Lecturer in Moral Science at the University of Cambridge. Copeland pointed out as Braithwaite's areas of expertise the philosophy of science and decision and game theory, all applied to moral philosophy. The transcript was first published by Copeland in (1999), and is available in the Turing Digital Archive. From what we see in it (2004 [1952]), Newman and Braithwaite were expected to play sort of referees in the presence of Jefferson and Turing, whose positions were polarized since *The Times*' June 1949 headlines. Braithwaite, in particular, was supposed to play the role of an anchor to the discussion.

Copeland wrote that it can be considered to have been the first recorded discussion on artificial intelligence (2004, p. 487). It is also, I shall add, the last record of Turing's defense of the idea of machine intelligence. About this event, as we have seen (§1.7), Turing wrote in a letter to his friend Norman Routledge: "[g]lad you enjoyed broadcast," and completed "Jefferson certainly was rather disappointing though" (2012 [*c.* early 1952]). Juxtaposed to this observation, he finished the letter with his syllogism in distress. Since the 1949 Manchester seminars when most likely Turing and Jefferson firstly met in person, over two years have passed in ongoing debate. Eventually, after this 1952 BBC roundtable, Turing seemingly got fed up. Now, through the lens of this last event, we shall see how this happened. Seeing this 1952 event as an important source, I shall present it in some detail. The number of interventions by each participant was balanced: Turing (15), Jefferson (16), Newman (16) and Braithwaite (13). I shall refer to them by the following notation, *e.g.*, 'T1' refers to Turing's first intervention, 'J16' to Jefferson's last one, and so on. I will present their discussion in a way that to my knowledge has never been presented, which is to follow the interventions, the rationale and point of view exposed by each participant. This shall prove worth it for the sake of this chronology of Turing's concept of

machine intelligence. Let us start with Braithwaite and Newman, the referees.

Richard Braithwaite's viewpoint

Braithwaite had started his (13) interventions by making questions to the experts (B1-B3). He made this introduction (B1) to warm up:

> *Braithwaite*: We're here today to discuss whether calculating machines can be said to think in any proper sense of the word. Thinking is ordinarily regarded as so much a speciality of man, and perhaps of other higher animals, that *the question may seem too absurd to be discussed*. But, of course, it all depends on what is to be included in thinking. The word is used to cover a multitude of different activities. What would you, Jefferson, as a physiologist, say [...]? (TURING et al., 2004 [1952], p. 494, emphasis added)

So Braithwaite opened by acknowledging the paradoxical aspect of the question "can machines think?," like Wolfe Mays did with his reference to the O.E.D. as I mentioned in the beginning of this chapter. (Turing himself, let us recall from §A.4.3, in the Step 3 of his 1950 paper, had done just the same. He acknowledged that the question was too meaningless to deserve discussion from the point of view of the common sense.) And yet, unlike Mays, Braithwaite's attitude was to keep it open for the discussion. He emphasized that the meaning of "thinking" shall in effect depend on what is to be included in its usage extension. This, independently of whether or not Braithwaite would agree with Turing's views, can be seen as a recognition of the intelligibility of Turing's proposal at face value.

Jefferson took over and the conversation went on. Then in his fourth contribution (B4), Braithwaite seems to have challenged Turing's general view on machine learning:

> *Braithwaite*: No-one has mentioned what seems to me the great difficulty about learning, since we've only discussed learning to solve a particular problem. But the most important part of human learning is learning from experience — not learning from one particular kind of experience, but being able to learn from experience in general. [...] The peculiarity of men and animals is that they have the power of adjusting themselves to almost all the features. [...Man] attends to what he is *interested in*. His interests are determined, by and large, by his appetites, desires, drives, instincts — all the things that together make up his 'springs of action'. (TURING et al., 2004 [1952], p. 497, no emphasis added)

After such strong observation, nonetheless, Braithwaite seems to have payed close attention to Turing and Newman's replies. He supported (B6) Turing's claim about the possibility of making a machine to find new concepts by a blind yet systematic combination of words. And then he went back to his anchoring role by provoking the others with interesting clarifying questions (B7-B9), up to his tenth intervention (B10) when he came back to his point about learning from experience in connection with emotions and appetites. He was actually disagreeing with Jefferson who had just stated (J14) the impossibility of machines to have some counterpart of human or animal emotions. Braithwaite considered that "[p]erhaps it will be impossible to build a

machine capable of learning in general from experience without incorporating in it an emotional apparatus." He suggested that, as "in humans tantrums frequently fulfil a definitive function," *viz.*, that of escaping too hard emotions or a too hostile environment, machines might need to incorporate them as well. Turing did not really follow it (T14).

Now, the next three interventions (B11-13) by Braithwaite were key. In his eleventh observation (B11), he streamlined the discussion, which for him was too much centered on whether machines could do everything that a man can do. He said: "[t]he point is, surely, whether they can do all that it is proper to call thinking." And completed: "[a]ll that it has got to do in order to think is to be able to solve, or to make a good attempt at solving, all the intellectual problems with which it might be confronted by [its] environment." At this point Newman jumped (N13) on Braithwaite, and thus let him make (B12) a very important intervention in connection to the Jefferson-Turing controversy. I shall call it henceforth Braithwaite's razor to the question whether machines can think, or *Braithwaite's razor* for short:

> *Newman*: But I thought it was you who said that a machine wouldn't be able to learn to adjust to its environment if it hadn't been provided with a set of appetites and all that went with them?
> *Braithwaite*: Yes, certainly. But the problems raised by a machine having appetites are not properly our concern today. It may be the case that it wouldn't be able to learn from experience without them; but we're only required to consider whether it would be able to learn at all — since I agree that being able to learn is an essential part of thinking.
> (TURING et al., 2004 [1952], p. 504)

Thus Braithwaite cut to the chase. *The question on whether a machine can have appetites, emotions, feelings etc. is not a direct requirement for thinking. It is rather being able to learn that is*. That question has been Jefferson's main point both in his Lister Oration, as we have seen in *The Times*' quotation of him (§A.4.1), and again in this 1952 BBC discussion, as we shall see shortly. Yet as Braithwaite posed most clearly, it should perhaps be a concern for those engaged in trying to make a machine to learn but it was no more than a red herring to their discussion. Moreover Braithwaite identified most spontaneously that there was alive in the discussion — as he said "I agree" — the view that *the capability of learning is a necessary property for thinking*. (And it can be traced back to Jefferson's third intervention.)

Braithwaite's last intervention (N13) was a question to Newman, as presented next.

Max Newman's viewpoint

Newman started his (16) interventions by responding to a most vivid description by Turing (T1-T2) of his imitation game or test for machine intelligence. Newman cheered it (N1) by saying that he would like to see the game to take place and try his hand at making up some of the questions. But he felt that it would take a long time to be feasible for the machine to do well in the test. He then asked Turing about it. He was then questioned directly by Jefferson (J2):

"what kind of things can [automatic calculating machines] do now?" Jefferson's intervention kept Newman emphasizing "mathematical computing" as the machines' "strongest line" and discussing (N3-N5) the machine's disposition for solving problems from playing chess to train scheduling (N3), and their potential for learning (N4) and remembering (N5). At some further point he wanted (N6) to bring the discussion back to capabilities that seemed feasible machines existing then at their time, and raised (N7) the question on whether and how the machine could invent a new (mathematical) concept. Turing replied and Newman asked him back (N8) about how efficient could the machine do it.

From then on Newman made the two rounds of, in my view, his most key contributions. In his ninth intervention, just before Turing had presented (T9) to Braithwaite his elaboration that thinking is an emotional concept, Newman went (N9) in support of Turing. And Newman gave this interesting ancient-church image about "what big computing machines actually do:"

> *Newman*: It is quite true that people are disappointed when they discover what the big computing machines actually do, which is just to add and multiply, and use the results to decide what further additions and multiplications to do. '*That's* not thinking', is the natural comment, but this is rather begging the question. If you go into one of the ancient churches in Ravenna you see some most beautiful pictures round the walls, but if you peer at them through binoculars you might say, 'Why, they aren't really pictures at all, but just a lot of little coloured stones with cement in between.' The machine's processes are mosaics of very simple standard parts, but the designs can be of great complexity, and it is not obvious where the limit is to the patterns of thought they could imitate.
> (TURING et al., 2004 [1952], p. 500, no emphasis added)

I take this to clearly side Newman with Turing with respect to have intelligence as an emotional concept. And he would emphasize it even more in this next intervention of his.

Braithwaite asked (B8) "how many stones are there in your mosaic?," meaning how many storage units were there in the Manchester machine, and at the same time posed to Jefferson the same question relative to cells in the brain. Conversation went on up to Jefferson saying (J11) that, in any case, "it is not, surely, just a question of size." He explained: "[t]here would be too much logic in your huge machine. It wouldn't be really like a human output of thought. To make it more like [a brain], a lot of the machine parts would have to be designed quite differently to give greater flexibility and more diverse possibilities of use." Turing disagreed (T12): "[i]t really is the size that matters in this case," and related it with finding one among several possible machines. This made Jefferson to come back to his point (J12) about the effects of external stimuli and some relation with the environment as required for (creative) thinking. Then uttered that he could not see how a machine could "as it were" say "Now Professor Newman or Mr. Turing, I don't like this programme at all that you've just put into me, in fact I'm not going to have anything to do with it." At this point, Newman made (N10) what is in my view this most interesting contribution which to my knowledge has not yet been identified. *Newman replied to Jefferson that even if he could make the machine to say so, Jefferson would not accept it*. In

effect, Newman's suggested that Turing's notion of intelligence as emotional was actually an identification of the problem of *confirmation bias* towards the view that machines *can't* think. To alert about this problem, Newman referred to his trouble in how to read the Bible:

> *Newman*: One difficulty about answering that is one that Turing has already mentioned [T9]. If someone says, 'Could a machine do this, e.g. could it say "I don't like the programme you have just put into me", and a programme for doing that very thing is duly produced, it is apt to have an artificial and ad hoc air, and appear to be more of a trick than a serious answer to the question. It is like those passages in the Bible, which worried me as a small boy, that say that such and such was done 'that the prophecy might be fulfilled which says' so and so. This always seemed to me a most unfair way of making sure that the prophecy came true. If I answer your question, Jefferson, by making a routine which simply caused the machine to say just the words 'Newman and Turing, I don't like your programme', you would certainly feel this was a rather childish trick, and not the answer to what you really wanted to know. But yet it's hard to pin down what you want. (TURING et al., 2004 [1952], p. 501-2)

In Newman's final interventions (N11-N16), he restated that he would like to see the discussion *not* to be focused on (N12) "what hypothetical future machines will do," and again emphasized the importance of how fast machines accomplish their intellectual tasks. He also engaged (N13) with Braithwaite to produce the latter's reply (B12) that I proposed above to call Braithwaite's razor. Newman took over from it and articulated a variant (N14) of what I have called (Cf. Introduction) *Newman's distinction*:

> *Newman*: There are really two questions that can be asked about machines and thinking, first, what do we require before we agree that the machine does *everything* that we call thinking? This is really what we have been talking about for most of the time; but there is also another interesting and important question: Where does the doubtful territory begin? What is the *nearest* thing to straight computing that the present machines perhaps can't do?
> (TURING et al., 2004 [1952], p. 504, no emphasis added)

Braithwaite then could not avoid asking (B13): "[a]nd what would your own answer be?" In coherence with what he had been defending throughout the discussion, *Newman formulated (N15) a test for machine intelligence* — which Copeland identified in (2004, p. 492-3) as "Newman's test" — that was coherent with his views. It required the machine to be able "to solve mathematical problems for which no method is known, in the way that men do; to find new methods." I interpret Newman's test as posing that the machine shall be able to prove non-obvious theorems. His last contribution (N16) was a low-profile answer to a rhetorical question by Jefferson (J15).

Geoffrey Jefferson's viewpoint

Jefferson started his (16) interventions by answering to Braithwaite's (B1) question about what are the most important elements involved in thinking. Jefferson gave a long, well-structured

answer (J1). He began by trying to deflate the importance of a definition: "I don't think that we need waste too much time on [a] definition of thinking since it will be hard to get beyond phrases in common usage, such as having ideas in the mind, cogitating, meditating, deliberating, solving problems or imagining." He added "I agree that we could no longer use the word 'thinking' in a sense that restricted it to man. No one would deny that many animals think, though in a very limited way. They lack insight." Jefferson shifted to the average person, who would in his view "be content to define thinking in very general terms such as revolving ideas in the mind." And then he proceeded to share with his fellow participants what his view was:

> *Jefferson*: One might say in the end that thinking was the general result of having a sufficiently complex nervous system. Very simple ones do not provide the creature with any problems that are not answered by simple reflex mechanisms. Thinking then becomes all the things that go on in one's brain, things that often end in an action but don't necessarily do so. I should say that it was the sum total of what the brain of man or animal does.
> (TURING et al., 2004 [1952], p. 494)

This (J1) is *Jefferson's 1952 definition of thinking*, and it can hardly be found explicitly like this neither in his Lister Oration (1949a) nor to my knowledge anywhere in his writings. It is in essence consistent with what Jefferson expressed in 1949. *For Jefferson*, I think we can thus summarize it further to connect the two pieces in his answer, *thinking is the sum total what the brain of man or animal does as a result of having a sufficiently complex nervous system.*

Jefferson's second and third interventions were to ask Newman: (J2) about what machines can do now; (J3) whether machines can "learn to do better with practice," which was for him "such an important point;" and (J4) "[h]ow long can a machine store information for?" This sequence of questions, while could be read as Jefferson's attempt to inform himself, in my view can be best read by observing his successive attempts to push the matter beyond the reach for machines after convincing positive answers given by Newman (N2-N5). His fifth contribution then was (J5) a speculation that machines do not learn like humans (under "frequent intervention by teachers, parental or otherwise"). It was promptly rebutted by Turing (T4) on the authority of his own initial experiments on machine learning. Jefferson tried to discredit Turing's response with the questions (J6) "[b]ut who was learning, you or the machine?" Turing replied (T5) "we both were," and extended his answer in detail hoping that machine learning could be accelerated by "a sort of snowball effect."

In his seventh contribution, following Braithwaite's (B4) note on the importance of learning from experience and appetites, Jefferson (J7) directly admonished Turing about suggesting that in a short time frame a replica of man would be artificially created. Although Turing had just guessed (T3) in reply to Newman (N1) that it would take at least 100 years for a machine to pass his test, (perhaps because in 1950 Turing had guessed 50 years instead) Jefferson said that since "the time of Descartes and Borelli on people have said that it would be only a matter of a few years, perhaps 3 or 4 or maybe 50" (p. 498).

From Jefferson's eighth intervention to the eleventh (J8-J11), he indicated lack of available knowledge on the workings of the human brain. At that point (J11), in comparing the machine with the brain, Jefferson stated his position on the machine-mindbrain analogy this way:

> *Jefferson*: You would need 20,000 or more of your machines to equate digits with nerve cells. But it is not, surely, just a question of size. There would be too much logic in your huge machine. It wouldn't be really like a human output of thought. To make it more like, a lot of the machine parts would have to be designed quite differently to give greater flexibility and more diverse possibilities of use. It's a very tall order indeed.
> (TURING et al., 2004 [1952], p. 501)

Turing replied sharply (T12): "[i]t really is the size that matters in this case." Not comfortable with Turing's reply, Jefferson proclaimed (J12) a stronger variant of Braithwaite's point about the importance of the environment and appetites (B4), framing it as a hard requirement not for learning but for thinking itself. He said: "[i]f we are really to get near to anything that can be truly called 'thinking' the effects of external stimuli cannot be missed out." And continued up to his comment that a machine should be able to utter its dislike about one or another program put into it. It was at this point that Newman alerted about the risk of engaging in confirmation bias, as we have seen (N10). In face of that, Jefferson replied (J13): "I want the machine to reject the problem because it offends it in some way." He then referred to people's varieties of tastes and located their sources in Mendelian inheritance, pointing that "[y]our machines have no genes, no pedigrees." In fact at this point Jefferson elevated his tone and, instead of bearing with his argument to Newman, he shifted to Turing and threw at him, I interpret, one of his subtle *ad hominem* arguments. He said (still in J13): "But I don't want to score debating points! We ought to make it clear that not even Turing thinks that all that he has to do is to put a skin on the machine and that it is alive!"

In his following interventions (J14), Jefferson restated the need for the machines to be influenced by emotions. Like in his 1949 Lister Oration, he emphasized the chemical aspect of the nervous system as "tremendously important." He said "man is essentially a chemical machine," and because machines are not, Jefferson added, they must be "'mentally' simple things" and "perform their tasks with an absence of distracting thoughts which is quite *inhuman*" (p. 502, no emphasis added). Braithwaite countered (B10): "I'm not sure that I agree." The conversation continued with Turing and Newman up to Jefferson's fifteenth intervention (J15), when he talked at length about our limited knowledge on the workings of the human brain. Eventually, at that point, he seems to have focused (J15) on a last battle to defend the limiting case (as in the famous quoting of his Lister Oration) of accepting that "machine equals brain." He involved Newman in contributing (N16) to a rhetorical question before his final remark (J16):

> *Jefferson*: [...] But, Newman, before we say 'not only does this machine think but also here in this machine we have an exact counterpart of the wiring and circuits of human nervous systems,' I ought to ask whether machines have

> been built or could be built which are as it were anatomically different, and yet produce the same work.
> *Newman*: The logical plan of all of them is rather similar, but certainly their anatomy, and I suppose you could say their physiology, varies a lot.
> *Jefferson*: Yes, that's what I imagined — we cannot then assume that any one of these electronic machines is a replica of part of a man's brain even though the result of its actions has to be conceded as thought.
> (TURING et al., 2004 [1952], p. 505-6)

In the ending of this last intervention (J16), Jefferson pushed to Newman *an instrumentalist view of the machine, as opposed to, say, Turing's realist view of the machine as an electronic brain*:

> *Jefferson*: [...] The real value of the machine to you is its end results, its performance, rather than that its plan reveals to us a model of our brains and nerves. Its usefulness lies in the fact that electricity travels along wires 2 or 3 million times faster than nerve impulses pass along nerves. [...] But that old slow coach, man, is the one with the ideas — or so I think. It would be fun some day, Turing, to listen to a discussion, say on the Fourth Programme, between two machines on why human beings think that they think!
> (TURING et al., 2004 [1952], p. 505-6)

Thus Jefferson, who was at that point discussing with Newman, all of a sudden shifted to Turing once more. No further comments were made after that, and the broadcast was ended.

Alan Turing's viewpoint

Turing started his (15) interventions taking over from Jefferson (J1). After having given his definition of thinking, Jefferson asked Turing whether he had "a mechanical definition" of it himself. This offered Turing an opportunity to add something to what he had wrote in his 1950 paper. Recall from §A.4.3 that he had dismissed the relevance of basing his discussion of the question "can machines think?" on commonsense definitions of the terms "machine" and "thinking." Now in 1952 Turing said (T1):

> *Jefferson*: [...] I should say that [thinking] was the sum total of what the brain of man or animal does. Turing, what do you think about it? Have you a mechanical definition?
> *Turing*: 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. The important thing is to try to draw a line between the properties of a brain, or of a man, that we want to discuss, and those that we don't. To take an extreme case, we are not interested in the fact that the brain has the consistency of cold porridge. We don't want to say 'This machine's quite hard, so it isn't a brain, and so it can't think.'
> (TURING et al., 2004 [1952], p. 494-5)

Turing saw no gain in agreeing on a definition. The "important thing," Turing remarked, was to select "the properties of a brain" that were worth of discussion. Turing thus addressed Jefferson's definition almost directly. Jefferson had just implied that thinking required the organic substrate

of the brain of man and animal to take place. Turing replied that, say, the consistency of the brain, harder or softer, was of no interest to the question under discussion.

Following up, Turing outlined a variant of his test (T1-T2). We shall analyze it (§3.3). For the sake of this chronology, it is important to keep track of a couple points. First, just like in his 1950 paper, Turing held that the original question "can machines can think?" should not be set aside or dismissed on account of being replaced by his test. He even said (T1) that "it would be better to avoid begging the question, and say that the machines that pass are (let's say) 'Grade A' machines." With this note, *Turing emphasized the relativity of his idea of a test for machine intelligence*. Second, as in the 1950 variant of his test, *Turing's 1952 presentation of the test takes conversational question answering as its intellectual task* likewise. But now Turing invoked (T1-T2) the scenario of a law court with the machine being interrogated by a jury. Newman, as we have seen, cheered Turing's 1952 test and said (N1) that he wanted to try his hand at questioning the machine, but he also suggested that it would be "a long time" from then for the machine to have a chance at such unrestricted conversational test. Turing agreed, and thereby (T3) postponed his prediction, now to come true in "at least 100 years" from the early 1950's.

Turing's next seven interventions (T4-T10) were on *machine learning*. Here is how it all started. Jefferson replied (J2) by asking Newman — not Turing — how well existing machines would stand up to Turing's test, and conversation went on with Newman describing a routine to make the machine to play chess and saying "I think this can fairly be called learning." Turing only intervened again when Jefferson questioned it for the second time by saying that the process looked too autonomous to be compared with human learning. Turing said (T4) that in his view of machine learning the education would be based on parental or teacher's intervention just like with human children. He related to have done initial experiments along those lines. Jefferson further questioned (J6) who was learning, him or the machine. Turing replied (T5) both, and proceeded in describing his view of the pace in which he expected the machine would learn by referring to an image of "snowball" learning: "the more things the machine has learnt the easier it ought to be for it to learn others." He added: "[i]n learning to do any particular thing it will probably also be learning to learn more efficiently." Turing then passed it to Braithwaite, who (B4) made his point as we have seen about the need for "learning from experience." Jefferson and Newman contributed to it, until Turing again took the word. He posed (T6) that in his view machines could "do something as advanced as finding a useful new concept." They could do it through random "combinations of words" together with a scoring scheme. What Turing described was a data-driven (statistical) approach for concept learning. ("Statistical learning" in general is the dominant approach today in artificial intelligence after several decades of dominance of the "knowledge-based" approach based on classical logics under influence of John McCarthy and others.) Newman asked (N8) about the time the machine would take for a feat such as learning a new concept. Turing answered (T7) it could take long but one should notice that the complexity of learning hard concepts could be broken up if his view of snowball learning was

effective. Braithwaite came then with a question (B6) concerning situations that may look similar to someone yet identifying their similarity in terms of mathematical structure can be challenging. Those would be hard if not impossible for a machine to identify in order to form analogies. In reply Turing came up (T8) himself with an analogy. He suggested that double negation in logics could be taught by association with crossing a road back and forth. Similar patterns of association could perhaps be taught to the machine by unveiling how learning is structured in the brain. Jefferson tried (J8) to conceal the idea of tying high-level (linguistic) descriptions to the structure of "cells and connecting fibres" in the brain. But Braithwaite wanted to hear more, and his intervention (B7) let Turing give a remarkable reply (T9). As identified by Copeland (2004, p. 491), Turing rephrased what he had already stated in his 1948 NPL report to make the point that intelligence is an emotional concept. He added "[f]rom this point of view one might be tempted to define thinking as consisting of 'those mental processes that we don't understand'." I discuss Turing's concept of thinking or intelligence in detail elsewhere (§2.3). The interesting news in this 1952 occasion is that it lent us to see Newman expressing his agreement with it, as we have seen (N9-N10). Jefferson replied (J9) dismissively, "[i]f you mean that we don't know the wiring in men, as it were, that is quite true. Turing said (T10) "[n]o, that isn't at all what I mean." He tried to further explain just what he had already said in the October 1949 seminar, had wrote in his 1950 paper in rebuttal to the argument from informality of behavior, and had said again in his May 1951 radio lecture (§A.4.5). Even if the wiring could be known in man as it is in the machine, one could still get surprised. That is, knowing the wiring of an entity may imply that its behavior is predictable, but that still does not mean that *in practice* one would be able to predict it.

Braithwaite's intervened (B8) and shifted the topic to the size of the computer's capacity in analogy with the number of cells in the brain, which led Turing to his next contributions (T11-T12). Turing stated that the number of digits that defines the storage capacity of the machine can be assumed to correspond to the number of cells in the brain. Jefferson countered (J11) about the comparison with the brain: "[y]ou would need 20,000 or more of your machines to equate digits with nerve cells." He continued "[b]ut it is not, surely, just a question of size," and added "[t]here would be too much logic in your huge machine." Now, the reader may recall that Turing had insisted in his May 1951 lecture (§A.4.5) that no "increase in the complexity of the computers" was needed to imitate a brain, but rather the use of "larger and larger computers" appropriately programmed. In 1952 Turing again emphasized (T12) "[i]t really is the size that matters in this case." This made Jefferson to go back (J12) to his demands about external stimuli and appetites, which led Newman to make that tough reply (N10) when he pointed out the problem of confirmation bias in Jefferson's position. Jefferson then (J13) turned to Turing (*ad hominem*) and implied that Turing's position included assigning infallibility to the machines, seen as an advantage over us humans. Turing could do nothing but invalidate it. He said (T13): "[c]computing machines aren't really infallible at all." But that only made Jefferson to return (J14) to his claim that "[a]t any rate," machines "are not influenced by the emotions." As we

have seen, Braithwaite disagreed (B10) and suggested that machines could be equipped with an "emotional apparatus." Turing did not really follow it (T14). He said he did not "envisage teaching the machine to throw temperamental scenes." Turing thought that some emotional behavior from the part of the machine would "likely occur as a sort of by-product of genuine teaching," but that, in his view, was to be curbed rather than encouraged. This adds strength to my interpretation (§1.5) that Turing's views shall not be related to transhumanism. In his final intervention (T15) Turing agreed with Newman (N12) that the "time factor is the one question which will involve all the real technical difficulty" in building the machines of the future.

Roundtable digest

In this account of the roundtable I have given visibility to the line of argumentation and focus of each participant. Braithwaite's argumentation, I interpret, was centered on learning from experience in general as a hard requirement for thinking and the relation of this with equipping the machine with an emotional apparatus. Newman wanted to center the discussion around mathematical applications that he thought were nearer to what machines could do back then. Jefferson presented a definition of thinking as the sum total of what the brain of man or animal does as a result of having a sufficiently complex nervous system. Informed about the new machines capabilities that by analogy matched human capabilities, such as learning and remembering, Jefferson has made his demands for a thinking machine a moving target. Although overall coherent with his 1949 Lister Oration, they seem to have been geared to set an impossible standard for the machines to achieve on the authority that they are *not* organic. Turing dismissed the need to agree on a definition of thinking, and held that the important thing was to select the properties of a brain that were worth of discussion. He gave a variant description of his test, and explained machine learning at length. Eventually, however, Turing also restated the problem that thinking seems to be an emotional concept.

A.4.7 The mind-machine controversy is faded out (Feb. 1952)

As reported by Paul Blum (2010, p. 57-8), from late January to early February 1952 Jefferson delivered a university extension course of three lectures (on 29 Jan., 5 and 12 Feb. 1952) he entitled "The workings of the human mind." The events were held by the "Extra-Mural Department," and the flyer read "Tickets for the course can be obtained, on payment of the fee, from the Director of Extra-Mural Studies, The University, Manchester, 13." Blum gathered the material from the Michael Polanyi Archive at the University of Chicago (2010). It also included notes taken by Polanyi at the third lecture. I highlight below an excerpt of Polanyi's minute notes (not quite readable) taken over Jefferson's lecture:

> Soul: We shan't know until everything the man does, is, etc. will have been explained mechanically. The sould [*sic*, should be read "soul"] the very last mechanically unexplainable thing which will remain.

> [...] The human mind can evolve something new, the machine can't. Though one is sometimes surprised at the results produced by the machine, the reaction is always that one might have thought of it in advance, it was just overlooked. Nothing basically new, inherent in the construction of the machine.
> (BLUM, 2010, p. 56-7)

As we have seen (§1.7), at this time Turing was dealing with the "sexual offenses" charges that were put on him. In any case, it seems that he did not attend Jefferson's lectures. Considering his note about Jefferson in his letter to Norman Routledge — "rather disappointing" —, Turing may not even have cared. He seems to just had let the controversy on machine intelligence to fade out. I propose to date the end of the Jefferson-Turing controversy to the BBC roundtable on 10 January 1952. It would have lasted precisely, accordingly, for two and half years.

A.4.8 "Chess" (*c.* late 1952)

It is also worth noting, finally, that less than one year after the BBC roundtable, English industrialist and former radar engineer B. Vivian Bowden published through Sir Isaac Pitman & Sons, Ltd. a collection, *Faster than thought: a symposium on digital computing machines* (the first edition appeared in London in 1 January 1953). This collection contained a text by Turing and others, as I now describe.

Bowden was the notable businessman behind the initiative of Ferranti, a Manchester-based technology company which produced commercial versions (Mark I, Mark II) of the Manchester Baby computing machine (§A.3.8). A comprehensive survey of the vision of Vivian Bowden in postwar Britain and of the historical context of *Faster than thought* is given by James Sumner (2014). Essentially, Sumner points out, Bowden laid the rhetorical groundwork to position the new machines as the natural outcome of a uniquely British technological trajectory. In the January 1953 collection, Bowden included a portrait of Lady Lovelace in the frontspiece, celebrated Charles Babbage, and sidetracked the notion of "electronic brains" to stick with the commercially safer "electronic digital computers." In his opening note, he wrote:

> During the last year or two most people must have heard of the remarkable devices often called "Electronic Brains"; every schoolboy knows that there are in existence some very complicated machines which are capable of astounding feats of arithmetic. This book contains descriptions of several of these monsters, an explanation of the way they work, and several essays describing how they can be used. We shall refer to them as "electronic digital computers," a name which describes them more accurately and is less contentious than the one which popular usage has favoured. (BOWDEN, 1953, p. vii)

The typescript of Turing's contributed text was entitled "Digital Computers applied to Games: Chess," but it was actually published just as "Chess" (2004 [1953]). Copeland determined that it was co-authored with Audrey Bates, Bowden himself, and Christopher Strachey. To my knowledge, the time when Turing wrote it with them is unknown, but must have been at some

point in 1952, probably late 1952. Thereby Turing just restated, now somewhat timidly, a few points he had already made before on machine intelligence. *As of* c. *late 1952, Turing restated somewhat timidly his beliefs that a machine can be made to play chess reasonably by learning from its experience, and it can be made also to perform well in his conversational question-answering test*. No change of thought at all is advanced by Turing in this occasion. However, because Turing then alluded to his test it has historical value and I shall take a moment to review it in what follows.

What does it mean to make a machine to play chess? Thus he opened this communication. Turing acknowledged that there are several possible meanings which might be assigned to that, and outlined a few:

> i) Could one make a machine which would obey the rules of chess, i.e. one which would play random legal moves, or which could tell one whether a given move is a legal one?
>
> ii) Could one make a machine which would solve chess problems, e.g. tell one whether, in a given position, white has a forced mate in three?
>
> iii) Could one make a machine which would play a reasonably good game of chess, i.e. which, confronted with an ordinary (that is, not particularly unusual) chess position, would after two or three minutes of calculation, indicate a passably good legal move?
>
> iv) Could one make a machine to play chess, and to improve its play, game by game, profiting from its experience? (TURING, 2004 [1953], p. 569)

While Turing outlined all of these senses having chess as the intellectual task to be addressed by the machine, note that he did not refer to his distinguishability criterion as he did in his 1948 NPL report. He then added to it two additional unrelated case scenarios by commenting sharply: "[t]o these we may add two further questions, unconnected with chess, which are likely to be on the tip of the reader's tongue:"

> v) Could one make a machine which would answer questions put to it, in such a way that it would not be possible to distinguish its answers from those of a man?
>
> vi) Could one make a machine which would have feelings like you and I do? (TURING, 2004 [1953], p. 569)

Now the distinguishability criterion appeared again, but not in connection with chess as intellectual task. So Turing posed side-by-side chess and conversational question answering as related challenges to test the capabilities of a machine, but identified the latter with his distinguishability criterion which is central in the imitation game. He informed that the problem to be considered in his text (2004 [1953]) was actually iii), or to make the machine to play a reasonably good game of chess. But in order to put this problem into perspective against the others, Turing wrote, he would give brief answers to each of them:

> To i) and ii) I should say 'This certainly can be done. If it has not been done already it is merely because there is something better to do.'
>
> Question iii) we are to consider in greater detail, but the short answer is 'Yes, but the better the standard of play required, the more complex will the machine be, and the more ingenious perhaps the designer.'
>
> To iv) and v) I should answer 'I believe so. I know of no really convincing argument to support this belief and certainly of none to disprove it.'
>
> To vi) I should say 'I shall never know, any more than I shall ever be quite certain that you feel as I do.' (TURING, 2004 [1953], p. 569)

I observe that Turing gathered options iv) and v) under the same interpretation or answer. He restated his belief that these are genuine possibilities. They could come true. A machine could be made to play chess and improve its play by learning from its experience. (As known, this milestone has been passed, indeed, at least since May 1997 as of the victory of machine *Deep Blue* over Garry Kasparov; cf., for example, Larry Greenemeier's story twenty years later in 2017). To win the Russian world champion, the machine was made to learn from its experience just in the way Turing had referred to.) A machine could also be made to answer questions put to it in a way indistinguishable from a man, and this is in fact a short version of Turing's 1950 imitation game or test. Also worth noting is that his answer to option vi) matches perfectly to his 1950 reply to the argument from consciousness, as we shall see elsewhere (§2.5).

A.5 Analytical summary

In this chapter I have outlined a chronology of the idea of machine intelligence in Turing's thought according to the core sources, primary and secondary. The reason why Turing engaged in an outrageous combination of words "machine" and "intelligence," his contenders could not accept, was that he did not himself commit to an ontological distinction between mind and machine when it comes to their intellectual capabilities. In this chronology, we have traced back Turing's moves in the pursuit of the extension and limits of mechanistic explanations of the nature of the human mind. To provide structure to our historical reasoning over Turing's thoughts, I have proposed three periodizations: the mostly foundational years (1936-1939), when Turing conjectured computing machines; the mostly experimental years (1939-1949), when Turing engaged with others in building computing machines; and the mostly dialogical years (1949-1952), when Turing got involved in an extensive debate about computing machines.

Foundational years (1936-1939). It was mostly in these years that Turing theorized about machines by drawing inspiration from the workings of the human mind. He thus established the mathematical foundations for the possibility of machine intelligence. In comparison with earlier thinkers such as Blaise Pascal, Gottfried Leibniz and Charles Babbage, he launched a disruptively new way to think about machines — not in terms of gears and wheels, but in terms of (axiomatic) mathematical definition and proof. I summarize as follows key specific developments that took place in this period of Turing's thought in connection with machine intelligence.

1936 May (Turing machines / §A.2.1) Turing showed the mathematical existence of a machine that could imitate some of the work done by our minds. Defined in terms of a finite number of configurations and programmed operations, the abstract machines that became known as Turing machines can perform any computation that can be described in an instruction table, or, in other words, any calculation that a human clerk can do with pencil and paper. When introducing the right way to think about (abstract) machines, Turing made extensive use of metaphors with human thinking, including words such as "memory," "aware," "states of mind" etc.. With no loss of generality, a Turing machine is an abstract special-purpose machine, in the sense that it will do the work described in a specific instruction table. And yet, Turing showed the mathematical existence of a particular Turing machine that can be said to be universal, as it can be made to imitate any special-purpose Turing machine. Turing's motivation to come up with his abstract machines was to address David Hilbert's *Entscheidungsproblem* or "decision problem," which lived in the realm of pure mathematics. This problem was addressed in 1931 by Kurt Gödel who made good progress and yet left it open. Turing gave a definitive answer to it. However, the technical results established by Gödel and himself still left technical gaps with potential connections to mechanistic explanations of mathematical activity and the human mind. Turing felt that they deserved further study which he would embraced next, in his doctoral thesis.

1938 Jun. (doctoral thesis / §A.2.2) Leaving interest or *initiative* aside, Turing regarded mathematical reasoning schematically as the exercise of a combination of two faculties: *intuition* and *ingenuity*. He referred to intuition, I interpreted, as the negative of consciousness. Ingenuity in turn would consist in aiding the intuition through suitable arrangements of propositions, and perhaps geometrical figures or drawings. If intuition, on the one hand, may not be replaced (in light of Gödel's incompleteness theorems), ingenuity, on the other hand, can. In his 1936 paper, one may interpret, ingenuity can arguably be replaced by a suitably programmed Turing machine. He introduced the concept of ordinal logics in his 1938 thesis as an attempt to further confine the significance of Gödel's incompleteness theorems, or in other words, to establish the strictest boundaries of the place of intuition in mathematics. An ordinal logic would afford a formal distinction between intuitive steps and mechanical steps in the proof of number-theoretic theorems. Turing succeeded in obtaining a completeness result. However, its deeper meanings in connection with the "metamathematical discussion" related to Gödel's argument was not clear. In connection with his completeness result, Turing suggested in correspondence with Max Newman a strategy to approximate "truth" by "provability," or to confine intuition in mathematics, as well as one pleases. The strategy was based on the use of "various machines," proof-checking, proof-finding and so on. And they would require only one actual machine to be realized — the universal machine. I posed that it is implied in Turing's comment that the role of the universal machine is interchangeable with the role of his ordinal logic. And that is the key connection between Turing's 1936 and 1938 works. After the publication of his thesis and during his wartime service Turing would learn that certainty or perfect accuracy, although required in a formal-logic system such as that of the Principia Mathematica, is not required for intelligence.

Experimental years (1939-1949). It was mostly in these years that Turing experimented with machines. Based on findings of Copeland's, there is a non-obvious way to describe Turing's work in the war. What he did in this time was actually to pioneer *machine intelligence*, with code breaking as the specific application or intellectual task that was programmed for special-purpose machines to do. In his experimental years Turing developed his notion of machine intelligence significantly further with respect to his foundational years. The heuristic techniques developed in wartime would be implemented in the postwar period into a universal computing machine, seen like a brain. In particular, Turing learned a different path towards machine intelligence in comparison to his work on mathematical logics during his foundational years. While solving codebreaking by using heuristic techniques, Turing observed that when it comes to the imitation of human thinking one may allow machines to make mistakes. I summarize as follows nine core sources to keep track of the new developments in Turing's thought during this period.

1938 Sep. – 1945 Jun. (Wartime service / §A.3.1). During his wartime service Turing developed with colleagues an efficiently computable heuristic technique for code breaking that they called "banburismus." Heuristics comprise shortcuts to prune the search space of a problem, and indeed turned to be a classical technique in modern artificial intelligence. Turing elaborated on machine intelligence still during the war, possibly as early as 1941, and before joining the NPL. In particular, he thought out how to program heuristics to make a machine to play chess.

1945 Dec. (NPL report / §A.3.2). Turing was recruited by the (British) National Physical Laboratory (NPL) to lead the design of the Automatic Computing Engine (ACE). This was meant to be an implementation of the universal Turing machine. But more than that, Turing said when joining the NPL, he was going to build "a brain." Turing's concept of machine intelligence is historically tied through the ACE to his concept of the universal (Turing) machine in the history of his ideas. Early from Turing's wartime service in 1941, and most clearly in late 1945 on, Turing worked with the game of chess as intellectual task and testbed for the development of his concept of machine intelligence. In that context, Turing made an occasional reference to the notion of allowing the machine to make mistakes.

1946 Oct. (Mountbatten's address / §A.3.3). Louis Mountbatten announced on 31 October 1946 the advent of the "electronic brain." It came out in Britain with some delay in relation to the announcement of the largest related American project, the ENIAC in Philadelphia. The repercussion around Mountbatten's speech can be viewed actually as an anticipation of the reception of Turing's ideas in British society and culture. He introduced the term "electronic brain," and alluded to "a revolution of the mind." Mountbatten's address has striking correlation with views on machine intelligence that were original from Turing, and there is multiple evidence that they came from Turing indeed. Mountbatten's address seemingly forced the NPL to move on from total secrecy. Following up in November 1946 Turing attended to two interviews about the ACE, portrayed as British's electronic brain. His bold views on machine intelligence appeared in the press side by side with the opposing views of Douglas Hartree, another British computer

pioneer. Turing was reported by *The Telegraph* to have said that he foresaw the time, "possibly in 30 years," when "it would be as easy to ask the machine a question as to ask a man." In another interview, asked if a machine would be able to play an average game of chess Turing is reported to have said that "we may be able to settle [it] experimentally in about 100 years time."

1946 c. Nov. (Letter to Ashby / §A.3.4). As of 1946-1947 (there is yet some uncertainty about the letter's precise date), we know of two aspects of machine intelligence, namely, (i) a baseline approach for machine learning and its limitations, and in connection with that, (ii) the extension and limits of a universal computing machine as an analogue model of the brain. Turing did not consider back then that the universal machine has to be necessarily uncritical when anything goes wrong and devoid of originality.

1947 Feb. (NPL lecture / §A.3.5). Early on in his 1947 NPL lecture, Turing said that digital computing machines such as the ACE are in fact practical versions of the universal machine. He addressed as a main topic of his talk the storage capacity of the ACE. For him, that would be a key property for the machine to be able to display some intelligence. Turing also thought as of 1947 that in order to show any sort of genuine intelligence, much larger storage capacities would be needed than were yet available. He estimated the memory capacity of the human brain to be of the order of 10^{10} binary digits, and that one might hope that some real progress on machine intelligence could be made with a storage capacity of 10^6 digits. That could be the case especially if one's investigations are confined to some limited field such as the game of chess. This is key source for Turing's rationale about choosing the game of chess as intellectual task and standard for thinking to experiment with. It was in his 1947 NPL lecture that Turing eventually returned to the objection to machine thinking based on Gödel's argument. He moved on from his formal-logic approach to it as of his 1938 doctoral thesis, and observed that certainty is not actually required for a machine to display intelligence.

1947 Spring (Meeting with Wiener / §A.3.6). Wiener has been reported to have said: "I defy you to describe a capacity of the human brain which I cannot duplicate with electronic devices." Wiener's position seems to be *aprioristic* and is laid positively towards strong mechanism. This differs from Turing's position in ways that will become clear from Turing's 1950 views on. In any case, both Turing and Wiener had been making bold claims relative to the analogy between electronic computing machines and the human brain. They met in the spring of 1947, and their encounter rendered Wiener to make some substantial citations of Turing. Thus, I interpreted, some cross-fertilization may have taken place between them, if not in terms of ideas themselves at least in regard to inspiration or stimulus. Yet another reason why the Wiener-Turing connection is important is that Jefferson's target in his 1949 Lister Oration was actually Wiener. But Turing in part shared views with Wiener, so much so to end up himself replying to Jefferson, firstly in the *The Times* and then in his famous 1950 paper.

1947 Jul. (Darwin's letter on Turing's leave / §A.3.7). Turing left the NPL in July 1947 for a sabbatical leave to never return. We have evidence on Turing's plans relative to machine

intelligence at that point from a memo by NPL director Charles Darwin. Turing wanted to extend his work on the machine still further towards the biological side. Until then the machine has been planned for work equivalent to that of the lower parts of the brain, and Turing wanted to see how much a machine can do for the higher ones. Turing wanted to know whether a machine could learn by experience, and this would be theoretical work to be done away from the NPL.

1948 Jun. (Manchester "Baby" / §A.3.8). The Manchester "Baby" computing machine was a project of Max Newman, who founded the Computing Laboratory. He recruited Turing as a Reader in the Mathematics Department and "Deputy Director" of the new lab, but also sought support from engineers Prof. F. C. Williams and Tom Kilburn to build the hardware equipment. This machine attracted Turing to the University of Manchester in the first place. It was in fact the first realization of a universal Turing machine, and the only actual general-purpose machine which Turing would be able to experiment with. As related by Newman later, Turing worked on the design of the sub-routines out of which the larger programs for the machine are built, and on more general problems of numerical analysis. It was also the living object that drew the attention of the press in Manchester and of Jefferson in particular, feeding his controversy with Turing.

1948 Summer (NPL report / §A.3.9). Turing's 1948 NPL report was a study on machine learning based on the analogy with the human brain in general, and with the (infant) human cortex in particular. It followed through with Turing's plans as of 1946 as he had written to Ross Ashby. Turing introduced a connectionist computation model that can be seen as a first version of what we call today neural networks. Turing briefly described a sort of dual approach to machine learning based on "discipline and initiative." Turing first presented a clear rationale about several intellectual fields, their pros and cons, and how well they suited for testing machine intelligence. Turing elaborated a new, far from obvious way to see how the concept of machine intelligence affects us. For him, intelligence or thinking is as an emotional concept. As of this time, summer of 1948, chess was Turing's preferred intellectual task to test for machine intelligence, and thus his initial experiments on the imitation game were performed by making a machine to play chess.

Dialogical years (1949-1952). In his foundational and experimental years, Turing interacted more with fellow mathematicians, electronic engineers and cyberneticians, and less with professional philosophers. Also, he seldom attended to interviews. Since June 1949, however, Turing expanded significantly his interlocution and got involved in a public controversy in the UK, most notably with Jefferson, on whether machines can think. I claim that this controversy is key for understanding Turing's imitation game or test for machine intelligence. From within it, Turing articulated his views on machine intelligence most clearly, dialogically.

1949 Jun. (Controversy started / §A.4.1). Turing's controversy with Jefferson can be dated to 11 June 1949, when Turing was quoted by *The Times* in reply to Jefferson's Lister Oration. In order to accept that *machine equals brain*, Jefferson demanded that machines should be able to write a sonnet or a concerto because of thoughts and emotions felt. Turing responded sharply and with irony to Jefferson's statements. He foresaw that the feats of electronic computing

machines back then was only the shadow of what was going to be. Some experience with them was needed, Turing said, before we could really know about their capabilities. He saw no reason why the machines could not enter any one of the fields normally covered by the human intellect, and eventually compete on equal terms. Turing also denied that one can draw the line about sonnets as Jefferson tried to do, and playfully said that a sonnet by the machine would be better be appreciated by another machine. Max Newman wrote a note to *The Times* and to the British Medical Journal (BMJ, where Jefferson's Lister Oration would appear), with a diplomatic tone. Newman proposed, as we have seen (cf. Introduction), a key distinction between the universal and the existential variants of the question on whether machines can think. He suggested that too much of the public attention was being given to the universal variant of the problem, and — just as Turing had himself proposed in his 1948 NPL report — referred to games such as "bridge, poker or chess" as a field in which to test the new machine capabilities. I presented evidence that Newman's view was closer to Turing's than it may appear on the surface. But an editorial of the BMJ portrayed Newman as sober and Turing as fabulous. The Jefferson-Turing controversy is an event in the social history of the philosophy of mind and the field that we call today artificial intelligence. It went through the years until January 1952 (cf. the follow-up events below), and will inform my analysis of the Turing test (§3).

1949 Oct. – Dec. (Manchester seminars / §A.4.2). On 27 October 1949 there was a crucial event towards Turing's 1950 paper in *Mind*. A seminar on "the mind and the computing machine" was held in the Department of Philosophy of the University of Manchester, co-chaired by Dorothy Emmet and Michael Polanyi. Based on the minute notes that have been preserved, I have described a few core elements from the point of view of this chronology. Like Turing, Newman considered that the mind/machine problem can be decided empirically and only empirically. Moreover, Newman abstracted the problem of producing the original Gödel paper as the question whether a machine can do anything radically new. So Newman shifted the discussion around Gödel's argument from *the mathematical objection* to *Lady Lovelace's objection*. Turing's 1949 reply in turn extended his previous (1947 and 1948) postwar replies to the mathematical objection. In particular, it was now correlated with Turing's 1938 view, in connection with the notions of intuition and ingenuity, proof-checking and proof-finding machines, and most notably with the idea of an ordinal logic as analogous to a universal Turing machine. For Turing, the universal machine is capable of turning itself into any other (proof-finding) machine. Besides, it is clear that Turing did not rule out the possibility for a machine to produce an argument such as Gödel's. Emmet posed that the vital difference between mind and machine seems to be that the latter is not conscious. Turing replied that the act of choosing a method to accomplish a task can be seen as the exercise of consciousness. Also implied in that reply, I interpreted, Turing thought that machines can afford self-referential operations. After that Turing would not make any other positive comment with explicit reference to consciousness. Polanyi made an appeal to his concept of the semantic function underlying the human mind, which in distinction to a machine, would not be fully specifiable. Turing's 1949 spontaneous reply to that was to say that

the mind is only said to be unspecifiable because it has *not yet been* specified. This discussion must have been the source of Turing's 1950 formulation of *the argument from informality of behavior*. Turing held that it would be impossible to find by observation alone the program inserted even into a simple machine, not to say the logic behind the human brain. Polanyi further replied that this should mean that one cannot decide logical problems by empirical methods. This observation can be seen, in effect, as intuition that the discussion was lacking a crucial element, namely, an *epistemological standard for thinking*. We know of another edition of the same Manchester seminar, which was held in December 1949 near Christmas Eve. It can be established that Turing's attention was drawn to the text of Jefferson's Lister Oration at some point from 10 June 1949 (on the occasion of his interview to *The Times*) on until the write up of his *Mind* paper (before the summer of 1950). It was most likely in January 1950, I have inferred. Besides, I have observed that since Turing's very first thoughts on machine intelligence during the war through Turing's 1945, 1947 and 1948 NPL reports and lectures up to the opening of the Manchester seminars in October 1949, Turing still had the game of chess as his preferred choice of intellectual task to illustrate and test for machine intelligence.

1950 c. early (*Mind* paper / §A.4.3). The structure and interpretation of the imitation game will be object of in-depth analysis elsewhere (§3). Here is a a general outlook of it with focus on what was new with respect to machine intelligence as a concept that Turing had actually been developing since 1936. Turing dismissed the problem of defining words "machine" and "thinking" according to the common sense back then as needed in order to address the question on whether machines can think. It was here that Turing changed to conversational question answering as his choice of intellectual task to be addressed in his imitation game or test. Turing's teachings showed that his 1950 argument is essentially a reflection upon a science, which adds strength to the view that Turing, as a philosopher, tried to keep his thoughts tied up to his science. Turing outlined his beliefs (§6 of his paper) relative to the question whether machines can think, offering two empirically testable predictions for the future. His technological prediction stated that in about fifty-years time it would be possible to program a machine (it would have enough storage capacity) to play the imitation game well. His social-and-cultural prediction stated that although the original question "[c]an machines think?" was too meaningless to deserve discussion, the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. Turing updated his five 1948 objections and included in 1950 four new objections. Three of them have dialogical connections with Polanyi and Jefferson. Turing gave as of 1950 a dominating image to explain the possibilities that he considered on the extension and limits of the machine-mindbrain analogy, which is the "skin of an onion" analogy. He also claimed that the problem of making a machine to play well the imitation game is mainly one of programming. In comparison to his 1948 NPL report, Turing pushed forward his views on the problem of how to educate a "child machine." He suggested a balanced approach between two sources of learning, one based on binary outcomes of situations the machine experiences, and the other based on an imperative, symbolic though

yet uncertain inference (not meant to "satisfy the most exacting logicians"). Turing's proposed methods seem to better materialize what he had called in 1948 "discipline and initiative." Overall, as I have developed before (§1.5), Turing's *Mind* paper is the occasion when he clearly assumed his role as prophet of the machines.

c. 1951 (BBC radio lecture / §A.4.4). In this radio lecture Turing wanted to challenge the conventional wisdom "You cannot make a machine to think for you." He developed an argument in three logical steps: first he extended his discussion of *the mathematical objection*; then proceeded to resume as well his discussion of *learning machines*; and then finished with his comment on the *possibility of (super)intelligent machines*. Again, commenting on Gödel's argument and variations, Turing pushed back the burden of proof towards human thinking. In doing so, he drew from an interesting example from the history of mathematics. He said if a machine takes too long to give the answer for a mathematical problem this cannot be regarded as very different from what happens with mathematicians, who have for instance worked for hundreds of years on the question as to whether Fermat's last theorem is true or not. He shared his beliefs and thus added to his 1950 "skin of an onion" analogy another statement to insist that he was not considering any *a priori* limits for the machines' intellectual capabilities relative to ours as human beings. In his reiterated discussion of learning machines, Turing emphasized that what he had in mind was to make the machine to "learn by experience." He discussed strategies for the machine education process in my view with no clear novelty with respect to 1950. But now he drew attention to two interesting related aspects. He said that faking the actual education of the machine is not allowed. I interpreted this as suggestive that, for Turing, the capability of learning for itself is a hard requirement for the machine to be considered intelligent. Moreover, Turing made the caveat that his specific suggestions about how best to educate the machine were preliminary, and should be seen as nothing more than an analysis of actual methods of education applied to human children as he knew them. Finally, as we have seen before, Turing said that intelligent machines are a genuine possibility, and that it seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. I have claimed that, although there was irony in Turing's words, he meant that the possibility of (super)intelligent machines is germane indeed.

1951 May (BBC radio lecture / §A.4.5). Turing suggested that he shared with a fellow mathematician the view that machines are brains, and that it is not altogether unreasonable to describe digital computers as brains. The man in the street, Turing held, is bound to be persuaded (*cp.* with his 1950 social-and-cultural prediction) of the intelligence of the new machines, as they are capable of intellectual feats of which he would be quite incapable. The scientists, on the other hand, tend to be contemptuous of that as a superstitious attitude. They seemed to be eluded, Turing considered, by how digital computers were actually used at their time, and by how they will probably mainly be used for many years to come. Turing wanted to acknowledge that he agreed with Lovelace's dictum to the extent that deals with how digital computers *are* used rather than how they *could* be used. He outlined core assumptions and properties of digital

computers: their universality, their suitability to be programmed to imitate any other machine, the meaning of the "mechanical" analogy with brains, and their storage capacity. Turing made then what is in my view the central point of his May 1951 radio lecture as an argument for describing digital computers as brains. It should be noticed, he said, that there is no need for any increase in the complexity of the computers used. If we try to imitate ever more complicated machines or brains we must use larger and larger computers to do it. For Turing, I emphasized, as long as enough storage capacity is provided, the bottleneck to make a digital computer to imitate a brain is in the software, not in the hardware. So the wisest ground on which to criticize his views, Turing remarked, is that although computers might be programmed to behave like brains, we do not at present know how this should be done. He considered to be leaving open the question as to whether we will or will not eventually succeed in finding such a program. He said to be personally inclined to believe that such a program will be found. Turing did have made this point earlier (in Step 4 of his 1950 paper), but now he gave it much more visibility within his argument. Turing also restated his 1950 belief that probably, at the end of the century, it would be possible to program a machine to perform well in what he described in 1950 as the imitation game. He also restated the format of the imitation game as a *viva-voce* examination. Turing wanted to acknowledge the issue that to behave like a brain seems to involve free will, but the behavior of a digital computer, when it has been programmed, is completely determined. On that note Turing saw two possibilities: free will is either an illusion or unobservable. In the latter case it would be forever reserved by dogma to human beings only, so there would be nothing one could do about it. In any case, Turing's conclusion is that a machine supposed to imitate a brain "must appear to behave as if it had free will." Turing suggested overall, I interpreted, that whatever its internal mechanisms are, including random elements or not, for a machine to be regarded as thinking it must be capable to surprise us. Turing was then ready to revisit the question of Lady Lovelace's dictum. Against the view of scientists like Douglas Hartree, Turing argued that there is no need to suppose that when we give orders to the machine we know exactly what we are doing, or what the consequences of these orders are going to be. Just like he did in the *c.* 1951 radio lecture, Turing once more emphasized that he is not committed with a specific approach for the machine education process, as long as it bears a close relation of that of teaching. Finally, as we have already seen (§1.5), Turing outlined his views on the possibility of (super)intelligent, now with a tone less loaded with irony. He said that the new danger of having machines superseding us in intelligence power is much closer, and it is remote but not astronomically remote, and certainly something which can give us anxiety. He then declared, as I interpreted, his positive humanistic hopes in the study of the extension and limits of machine intelligence, and his dislike to the dismissal of the ontological distinction between the natural and the artificial, which I interpreted as a dislike to transhumanistic projects. Overall, I hold, Turing thought that (super)intelligent machines are a true possibility.

1952 Jan. (BBC roundtable / §A.4.6). This is the last record of Turing's defense of the idea of machine intelligence. In my account of the roundtable I have given visibility to the line

of argumentation and focus of each participant, summarized as follows.

Braithwaite. He opened the discussion by acknowledging the paradoxical aspect of the question "can machines think?" But his attitude was to keep it open for the discussion. He emphasized that the meaning of "thinking" shall in effect depend on what is to be included in its usage extension. This, I interpreted, can be seen as a recognition of the intelligibility of Turing's proposal at face value. Braithwaite made an important observation about learning from experience in general as opposed to learning to solve a particular problem. In man it would be determined by his appetites, desires, drives, instincts, all that together make up his 'springs of action.' Braithwaite played an anchoring role by provoking the others with interesting clarifying questions. He disagreed with Jefferson about the impossibility of machines to have some counterpart of human or animal emotions. He also thought that all that a machine has got to do in order to think is to be able to solve, or to make a good attempt at solving, all the intellectual problems with which it might be confronted by its environment. Braithwaite made a very important intervention in connection to the Jefferson-Turing controversy, which I called Braithwaite's razor. The question on whether a machine can have appetites, emotions, feelings etc. is not a direct requirement for thinking. It is rather being able to learn that is. That question had been Jefferson's main point in his 1949 Lister Oration and in this 1952 BBC discussion. For Braithwaite, it should perhaps be a concern for the sake of making a machine to learn but it was no more than a red herring to their discussion. He also identified that there was alive in the discussion the view that the capability of learning is a necessary property for thinking.

Newman. He first expressed his wish to try his hand at questioning a machine in Turing's test but also by giving his impression that it would take a long time for such a unrestricted conversation to be feasible for a machine. He described mathematical applications as their strongest line. By doing so he stressed their potential for learning and remembering. At some further point he wanted to bring the discussion back to capabilities that seemed feasible to the machines then existing. He raised the question on whether and how the machine could invent a new (mathematical) concept. Turing replied and Newman then asked back about how efficient could the machine do it. Newman agreed with Turing on the emotional aspect of the notion of thinking, and added to it by presenting an image about mosaics in an ancient church. The pictures can only be seen in proper perspective out of very simple colored stones. He also implied, I interpreted, that Jefferson's position was beset with confirmation bias towards the view that machines *can't* think. To alert about this problem, Newman referred to his trouble when he was a child in how to read the Bible. Newman restated that he would like to see the discussion *not* to be focused on what hypothetical future machines will do, and again emphasized the importance of how fast machines accomplish their intellectual tasks. He restated a distinction he first outlined in 1949, which I have called Newman's distinction, between a universal and an existential variant of the question on whether machines can think. Instead of discussing whether machines could do *everything* that we call thinking, they should discuss whether they can do the *nearest* thing to straight computing that we may call thinking. This would also be an interesting and important

question. Finally, he formulated a test for machine intelligence that was coherent with his views. It required that the machine shall be able to prove non-obvious theorems.

Jefferson. He first offered a definition of thinking, which can hardly be found explicitly neither in his 1949 Lister Oration nor to my knowledge anywhere in his writings. For Jefferson, thinking is the sum total of what the brain of man or animal does as a result of having a sufficiently complex nervous system. He questioned what machines could do then in the early 1950's, whether they can learn to do better with practice, and how long a machine could store information. He speculated that machines do not learn like humans, under frequent intervention by teachers, parental or otherwise. Jefferson directly admonished Turing about suggesting that in a short time frame a replica of man would be artificially created. He said that since the time of Descartes and Borelli on people have said that it would be only a matter of a few years. He indicated lack of available knowledge on the workings of the human brain. And yet, about the machine-mindbrain analogy, Jefferson stated that it is not, surely, just a question of size. There would be too much logic in your huge machine. It wouldn't be really like a human output of thought. He proclaimed a stronger variant of Braithwaite's point about the importance of the environment and appetites, framing it as a hard requirement not for learning but for thinking itself. Also, a machine should, he emphasized, be able to utter its dislike about one or another program put into it. Newman replied that even if he could make the machine to say so, Jefferson would not accept it. Jefferson then said he wanted the machine to reject the problem because it offends it in some way. He referred to people's varieties of tastes and located their sources in Mendelian inheritance, but machines in turn, he said, have no genes, no pedigrees. He also elevated his tone and, instead of bearing with his argument to Newman, he shifted to Turing and threw at him, I interpret, one of his subtle *ad hominem* arguments. Jefferson restated the need for the machines to be influenced by emotions. Like in his 1949 Lister Oration, Jefferson insisted on what he took to be a tremendously important point, that man is essentially a chemical machine, and because machines are not, they must be 'mentally' simple things' which perform their task in a way that is quite inhuman. Finally, Jefferson pushed to Newman an *instrumentalist view* of the machine, as opposed to, say, Turing's realist view of the machine as an electronic brain. He finished by referring nominally to Turing and saying that it would be fun some day to listen to a discussion on BBC between two machines on why human beings think that they think.

Turing. He first remarked that, rather than agreeing on a definition, the important thing was to select the properties of a brain that were worth of discussion. For Turing, the consistency of the brain, for example, was of no interest to the question on whether machines can think. Just like in his 1950 paper, Turing held that this question should not be set aside or dismissed on account of being replaced by his test. He suggested that machines that pass his test could be said 'Grade A' machines. He thus emphasized, I interpreted, the relativity of his idea of a test for machine intelligence. Also like in the 1950 variant of his test, Turing's 1952 presentation of the test takes conversational question answering as its intellectual task. But now Turing invoked the scenario of a law court with the machine being interrogated by a jury. In regard to his 1950

technological prediction on when a machine would perform well in his test, Turing conceded to Newman that it would take long(er), now to come true in "at least 100 years" from the early 1950's. Turing gave a lot of thought on machine learning. He insisted that in his view the machine education should be based on parental or teacher's intervention just like with human children. He also expected the machine to improve its learning rate (and learn how to learn) like a "snowball." For him, also, machines could do advanced things such as finding a useful new concept. It could take long but one should observe that the complexity of learning hard concepts could be broken up if his view of snowball learning was effective. Turing thought that machines could be made to spot an analogy. He gave an alternative phrasing to his 1948 notion of intelligence or thinking as an emotional concept, and said that from this point of view one might be tempted to define thinking as consisting of 'those mental processes that we don't understand.' He tried to explain to Jefferson what he had already stated in the October 1949 seminar, in his 1950 paper and in his May 1951 radio lecture, that knowing the wiring of an entity may imply that its behavior is predictable, but that still does not mean that in practice one would be able to predict it. About the analogy of digits and cells in machine and brain, as well as on the possibility of the former imitating the latter, Turing kept his May 1951 position and insisted that it really is the size that matters in this case. On the topic of human and machine fallibility, he held that computing machines aren't really infallible at all. Besides, he did not envisage teaching the machine to throw temperamental scenes. Emotional behavior, he thought, was to be curbed rather than encouraged. I interpreted that as more evidence unfavorable to associating Turing's views to transhumanism. Turing thought that the time factor is the one question which will involve all the real technical difficulty in building the machines of the future.

Roundtable digest. Braithwaite's argumentation, I interpreted, was centered on learning from experience in general as a hard requirement for thinking and the relation of this with equipping the machine with an emotional apparatus. Newman wanted to center the discussion around mathematical applications that he thought were nearer to what machines could do back then. Jefferson presented a definition of thinking as the sum total of what the brain of man or animal does as a result of having a sufficiently complex nervous system. Informed about the new machines capabilities that by analogy matched human capabilities, such as learning and remembering, Jefferson has made his demands for a thinking machine a moving target. Although overall coherent with his 1949 Lister Oration, they seem to have been geared to set an impossible standard for the machines to achieve on the authority that they are *not* organic. Turing dismissed the need to agree on a definition of thinking, and held that the important thing was to select the properties of a brain that were worth of discussion. He gave a variant description of his test, and explained machine learning at length. But eventually, Turing also restated the problem that thinking seems to be an emotional concept. About Jefferson's participation in it, Turing wrote some weeks later that it "certainly was rather disappointing."

1952 Feb. (Controversy ended / §A.4.7). From late January to early February 1952 Jefferson delivered a university extension course of three lectures on "The workings of the

human mind." Polanyi attended them, but the evidence is suggestive that Turing did not. At this time Turing was dealing with the "sexual offenses" charges that were put on him. He seems to just had let the controversy on machine intelligence to fade out. I propose to date the end of the Jefferson-Turing controversy to the BBC roundtable on 10 January 1952.

1952 c. *late 1952* (Text on chess / §A.4.8). As of *c.* late 1952, Turing restated somewhat timidly his beliefs that a machine can be made to play chess reasonably and learn from its experience, and it can be made also to perform well in his conversational question answering test.

A.6 Chapter acknowledgements

I thank João Cortese for the suggestion of word "dialogical" instead of "dialectical" to categorize Turing's argumentation and overall debating experience. While Turing's argumentation is in fact dialectical in the Socratic sense, I found the word "dialogical" to fit in very well to refer to Turing's (then) dialogical years (1949-1952) in general. I benefited from a discussion with João Cortese and Prof. Ricardo Terra on the historiographical problem of distinguishing public intellectual history from an author's individual intellectual history — in this dissertation, what I have been referring to as a chronology of Turing's thought. This has been a specially fruitful methodological topic for the study of Turing's ideas because of the state secrecy involved in much of his works. I also thank Turing historians for their curation and digital dissemination of documents that before were only available in the Turing archives at King's College, Cambridge.